AJTDE Volume 5, Number 2, June 2017

Table of Contents

A Time for Renewal

Editorial

Mark A Gregory
RMIT University

Abstract: The Telecommunications Association has commenced the second phase of the renewal process that started in 2013. As part of the first phase of this renewal process a key decision was to relaunch the Telecommunications Journal of Australia as the Australian Journal of Telecommunications and the Digital Economy. In addition to the Journal, the Telecommunications Association holds the Henry Sutton and Charles Todd Orations each year in Melbourne and Sydney respectively. As the second phase progresses, the Telecommunications Association will launch a new brand, update the Association's website and host the first of what should become an annual two-day telecommunications forum. The first event is to be held in Melbourne in November. For the Journal, key milestones have now been achieved, including being added to the SCOPUS list of indexed Journals and the Australian Research Council's Excellence in Research for Australia, which is the national research evaluation framework.

In This Issue

In this issue, the *Journal* includes topical articles that cover Australian telecommunications, historical events and an article on the state of the U.S. telecommunications industry.

The role of regulation in preventing Wi-Fi over-congestion in densely populated areas highlights the need to ensure that Wi-Fi infrastructure installation is managed to prevent a loss of performance caused by interference.

Implementation of PCC-OFDM on a software-defined radio testbed provides a description of the development of a testbed used in the development of a software-defined radio implementation of polynomial cancellation coded orthogonal frequency division multiplexing on a field programmable gate array based hardware platform.

http://doi.org/10.18080/ajtde.v5n2.114

Interference to Telephone Lines introduces a historical paper from 1936 that explores the effects of electrification of country Tasmania and the increasing interference to telecommunication circuits by high voltage power lines installed in close proximity.

An Evaluation and Enhancement of a Novel IoT Joining Protocol provides a description of a novel and enhanced Internet of Things to Wi-Fi network joining protocol.

Utilisation of DANGER and PAMP signals to detect a MANET Packet Storage Time Attack presents an approach using Artificial Immune System based Danger signals and a Pathogen Associated Molecular Pattern signal to identify a Packet Storage Time routing attack in Mobile Ad hoc Networks.

Fact of Fraud? describes the rise of the Repetitive Strain Injury phenomena in Telecom Australia over the period 1983-86.

U.S. Telco Industry History as a Prologue to its Future considers historical events as a means to highlight the competitive tension between the forces of competition and concentration and the latest battlefield - net neutrality.

Household bandwidth and the 'need for speed': Evaluating the impact of active queue management for home internet traffic aims to contribute to the policy debate on bandwidth needs by considering more closely what happens in household networks.

Social Network Behaviour Inferred from O-D Pair Traffic utilises anonymized Internet Trace Datasets obtained from the Center for Applied Internet Data Analysis to identify and estimate characteristics of the underlying social network from the overall traffic.

Making ICT Decommissioning Sexy! Challenges and Opportunities provides an insight into the challenges of decommissioning decision making and uses a framework and case study analysis as a means to improve the timeliness and financial motivations.

A Time for Renewal

The Telecommunications Association, publisher of this Journal, has commenced the second phase of a renewal process that began in 2013. The next steps include launching a new brand, updating the Association's website and hosting the first of what should become an annual two-day telecommunications forum.

The Telecommunications Forum 2017 is to be held in Melbourne in November with a focus on telecommunications and security. The Australian Government's telecommunication sector security reforms program has the potential to send a minor tremor through the industry later this year as legislation is passed that could start a major security shakeup within the industry in response to growing international cyber threats.

http://doi.org/10.18080/ajtde.v5n2.114

For the Journal, the past year has seen key milestones achieved. The Journal has been added to the SCOPUS list of indexed Journals and over coming years the Journal will build credibility as papers are added and the Journal's impact factor increases. It is difficult, of course, for public policy related papers about Australian and international telecommunications markets to attract a large number of citations, and it is for this reason the Journal continues to seek high quality papers that present new and novel research related to telecommunications and the digital economy.

The Journal has been included in the Australian Research Council's Excellence in Research for Australia, which is the national research evaluation framework. This is a major outcome for the Journal as it highlights the perceived value of the papers published in the Journal.

Over coming months, the Journal website will be updated and additional information added to assist authors through the publishing process.

Looking Forward

The key themes for 2017 will be *International Telecommunications Legislation and Regulations* and *International Mobile Cellular Regulation and Competition*. As the global digital economy evolves it is timely to consider the different telecommunications markets and how each is coping with the transition to next generation networks – the 'gigabit race' – and how competition is being fostered with the market. Mobile cellular continues to be an expensive consumer product and for many nations the promise of a competitive mobile cellular market has not eventuated due to the inherent advantages enjoyed by incumbent telecommunication companies during the deregulation years.

Papers are invited for upcoming issues and with your contributions the Journal will continue to provide the readership with exciting and informative papers covering a range of local and international topics. The Editorial Board values input from our readership so please let us know what themes you would like to see in the coming year.

All papers related to telecommunications and the digital economy are welcome and will be considered for publication after a peer-review process.

Mark A Gregory

http://doi.org/10.18080/ajtde.v5n2.114

The role of regulation in preventing Wi-Fi over-congestion in densely populated areas

Frank den Hartog
DoVes Research

Jan de Nijs
TNO

Abstract: Given the ever increasing number of Wi-Fi devices in use by the public, the progressing urbanisation, and the current attempts by the industry to improve Wi-Fi system performance, we here analyse the case of apartment blocks with residents increasingly suffering from Wi-Fi over-congestion. Here, individuals use private Wi-Fi networks in an "in house" environment to achieve cordless connectivity to the Internet. We show that Wi-Fi in apartment blocks is a true commons and, therefore, over-congestion can only be avoided by having the individual access point (AP) operators collaborating with each other. We found that such collaboration is not inhibited by current regulation, but neither can it be enforced. However, as AP operators will most likely enter collaboration voluntarily, further regulation is not deemed necessary.

Keywords: Wi-Fi, commons, congestion, regulation, urbanisation.

Introduction

Over the last decade, the use of portable devices has spectacularly increased. Wi-Fi connectivity has proven to be a primary asset of these devices. The Wi-Fi Alliance estimates that more than 8 billion Wi-Fi devices are currently in use around the world (Wi-Fi Alliance, 2017). Gartner expects that this number is going to increase much further, up to 21 billion devices, including not only tablets and smart phones, but ever more other types of embedded devices (called “things”) such as smart wearables and smart meters (Plummer et al, 2016). Many of these things will have to operate within homes. By 2020, according to Gartner, 20% of homes in the US will be connected homes containing more than 25 things accessing the Internet (Plummer et al, 2016).

Also technologies other than Wi-Fi, such as Zigbee and Bluetooth, are increasingly used to connect these devices with each other and the Internet. These technologies have in common

http://doi.org/10.18080/ajtde.v5n2.93

that they use so-called unlicensed or Class Licence spectrum: in most cases, individuals do not need to obtain a special licence before they can operate such wireless devices or access points (APs). The downside of this arrangement is that other operators may use these frequencies as well, also within the range of the network of the first operator, and as such introduce mutual interference, negatively influencing the performance of the networks involved.

This problem is well known to the industry, but so far it has not stopped the market surge of Wi-Fi and Bluetooth in particular. This is largely because the problem is not felt as long as the density of networks is low, i.e. as long as there are not too many networks operating concurrently within the same frequency band and within the same general location. This has certainly been the case for Wi-Fi and Bluetooth since they were invented in 1999, due to their relatively short range of about 10–20 metres indoor, which is largely determined by transmission power being capped by regulation. However, this relatively undisturbed existence is rapidly changing, not only because the number of wireless devices per household is increasing, but also because of the ever continuing urbanisation and densification of the cities.

In 2014, 54% of the world's population was urban, a number which is expected to rise to 66% by 2050 (United Nations, 2015). In Australia, 22 million people will be living in cities by 2020, an increase of 38% since 2000. In addition, most cities are executing policies of urban consolidation. Melbourne, for instance, aims to reduce the proportion of new development occurring at low densities on Melbourne's fringe from about 60% of annual construction to 40% by redirecting new development to defined areas of established inner and middle-ring suburbs (Melbourne 2030, 2002). As a result, more and more Australian residents are living in apartment blocks. Indeed, in the first three months of 2016, 29,987 apartments commenced construction in Australia, a number that for the first time in history overshadowed the figure registered for houses (25,122) (ABS, 2016).

Although there is plenty of anecdotal evidence that many users are already experiencing Wi-Fi congestion problems in densely built-up areas, and suffer from serious performance degradation as a result (Ozyagci et al, 2013), surprisingly little research has been done so far to quantify the pervasiveness and severity of the problem (De Vries et al, 2013), let alone to solve it. The European Horizon 2020 project "Wi-5" (Wi-5, 2015) recently produced the first quantitative evidence of this issue (Den Hartog, Popescu et al, 2016). It also proposes an architecture based on an integrated and coordinated set of smart solutions aimed at the efficient reduction of interference between neighbouring APs (Bouhafs, 2015). But as other authors have concluded before (De Vries et al, 2013), a solution can only be successful in the market if the issue is treated as a joint engineering, regulatory, and economic problem.

http://doi.org/10.18080/ajtde.v5n2.93

In this article, we focus on the regulatory aspects of Wi-Fi congestion in apartment blocks. Here, individuals use private Wi-Fi networks to achieve cordless connectivity to the Internet in an "in house" environment, instead of receiving a public Wi-Fi service from a commercial Service Provider. In the following section we describe the problem at hand in economic as well as technical terms. We then show that, from an economic perspective, we are dealing with a typical example of the Tragedy of the Commons. As a consequence, a solution can only be formulated in terms of some form of coordination among the APs' operators. We first give a simple example of how this could work in practice, and then postulate a generic business model for the more complex cases. We then treat the regulatory aspects of this business model in more depth, especially from the perspective of European and Australian spectrum access and antitrust laws.

Congestion of Class Licence spectrum

Spectrum commons

Much literature has been written on the question if spectrum for wireless communication is a common or a public good, or even a private good or a club good. Currently, most of the radio spectrum is managed and regulated by governments in a command-and-control fashion, i.e. spectrum is treated as a public good. A brief history on how this has evolved over the decades is provided by Peter Anker (Anker, 2017). In practice this means that national regulators are the centralised authorities for spectrum allocation and usage decisions. The usage is often set to be exclusive: each band is licensed to a single provider, thus maintaining interference-free communication. To enable international business and operation, national governments try, not always successfully, to standardise their policies in the ITU-R (International Telecommunication Union – Radiocommunications sector). In Australia, spectrum is regulated by the Australian Communications and Media Authority (ACMA). The Australian Government and ACMA follow the recommendations of the ITU as far as practical but not necessarily exclusively.

For the first time in 1947, the ITU-R set aside a number of frequency bands for the use of industrial, scientific and medical (ISM) purposes other than telecommunications. This allows devices that use RF for the purpose of heating (e.g. microwave ovens), medical diagnostics, etc. to “leak” radio waves as long as specified limits on transmit power are not exceeded. Operators of such devices do not need to acquire a licence. Well-known worldwide ISM bands include the frequency ranges of 2.4-2.5 GHz and 5.725 GHz - 5.875 GHz. As these ISM devices typically only emit and do not (intend to) receive signals, the ISM spectrum is rather a dump than a commons.

http://doi.org/10.18080/ajtde.v5n2.93

This changed in the 1980's, when communication systems were also allowed to use these ISM bands, under strict limitations such as the use of spread spectrum and listen-before-talk technology. In the US, this type of use is regulated by the Federal Communications Commission (FCC) in the Code of Federal Regulations, Title 47, Part 15 (GPO, 2016), and is called "unlicensed". As a consequence, the frequency bands used by these devices became popularly known as "unlicensed spectrum". In contrast to the original use of the ISM band, the unlicensed spectrum is a commons which every communication device operator can use at its own discretion, whilst having to accept the interference caused by other devices. The regulatory and economic aspects of these commons have been discussed and analysed by many authors already, also in this journal, especially in the context of public Wi-Fi deployment (De Vries et al, 2013; Weiser & Hatfield, 2005; Speta, 2008; Reed & Lansford, 2013; Potts, 2014; Goggin, 2014; Lambert et al, 2014).

Interestingly though, when discussing public Wi-Fi, Jason Potts (Potts, 2014) concludes that Wi-Fi is not a commons but a club good (Buchanan, 1965), as it is excludable (access can be denied) and non-rivalrous (access by one does not exclude access by others). The latter, though, is a consequence of the fact that the roll-out of public Wi-Fi by competing instances is regulated by governments by means of issuing Carrier Licences. In Australia, as in other countries, under the Telecommunications Act 1997 (Telecommunications Act, 2016), every owner of a Wireless Local Area Network (WLAN) network unit must have a Carrier Licence if the network unit is used to supply a carriage service to the public, i.e. to people outside the immediate circle of the network unit owner. This is not how Wi-Fi-enabled home gateways and other customer premises equipment (CPE) are typically used, and thus this rule does not apply.

Nevertheless, Australian law requires that all radio communications transmitters must be operated under the authority and in accordance with the requirements of a radio communications licence. This includes Wi-Fi devices and other devices operating in the ISM bands. So, technically speaking, unlicensed bands do not exist in Australia. For the majority of low-powered transmitters, the relevant licence is a so-called Class Licence, in this case the Radiocommunications (Low Interference Potential Devices) Class Licence 2015 (ACMA, 2016). This is a licence for which an operator does not need to apply and for which no licence fees are payable. Under a Class Licence, all users share the same spectrum segment and are subject to the same conditions. And as Wi-Fi-enabled CPE operation is Carrier Licence exempt, the Class Licence spectrum in densely populated residential areas such as modern apartment blocks, can be treated as a spectrum commons: it is non-excludable (resident A cannot forbid resident B to have a Wi-Fi home network), and it is rivalrous (if resident A uses

http://doi.org/10.18080/ajtde.v5n2.93

his Wi-Fi, it will have a negative effect on the performance of the network of resident B due to interference and congestion).

Wi-Fi

Today, IEEE 802.11-2016 (IEEE 802.11, 2016) is the main standard for WLANs, and the main implementations of IEEE 802.11 currently on the market are known as Wi-Fi. It includes all amendments to the original IEEE 802.11 standard from 1997, including IEEE 802.11b (".11b"), IEEE 802.11g (".11g"), IEEE 802.11n (".11n"), and IEEE 802.11ac (".11ac"), which are now commonly found in many households. Versions .11b and .11g operate in the 2.4 GHz ISM band only. Depending on the nationally specified width of the ISM band, these standards divide the band into 11-14 heavily overlapping channels. Every .11b or .11g WLAN is working on one of these channels, which is manually configured or auto-configured by a frequency selection algorithm in the AP. Only 3 channels can be chosen such that they do not overlap, e.g. channel 1, 6, and 11, or 1, 7, and 14.

The newer extensions of the standard, .11n and .11ac, also specify operation in the so-called 5 GHz band. This band stretches from 5.150-5.875 GHz. Technically speaking, it is not a single band though. The upper part of the band (5.725-5.875 GHz) is an ISM band, whereas the lower part is divided up into a multitude of sub-bands for which different rules apply regarding the maximum transmit power, indoor or outdoor use, how to implement channel bonding, and the obligation of implementing radar signal avoidance mechanisms such as Dynamic Frequency Selection (DFS) and Transmit Power Control (TPC). These regulations also differ slightly from country to country. As a result, most vendors have implemented only the lower four channels, i.e. a single band for which the requirements are everywhere the same.

Wi-Fi congestion and over-congestion

In 2013, Jean Pierre de Vries and co-authors published a comprehensive overview of the literature on Wi-Fi congestion, and concluded that excessive load is quite rarely observed, and very seldom well documented (De Vries et al, 2013). Often, when authors want to indicate the existence of congestion, figures similar to Figure 1 are shown. Figure 1 shows how many different APs, operating at different channels, are concurrently using the 2.4 GHz band, as observed in January 2017 with the tool inSSIDer (Home Version, version 3.0.7.48, MetaGeek), in a large apartment block in Belconnen (Canberra), Australia. In this case, 69 APs were found to be using overlapping channels. The measuring laptop was attached to the NetComm Wireless Service Set Identifier (SSID).

http://doi.org/10.18080/ajtde.v5n2.93

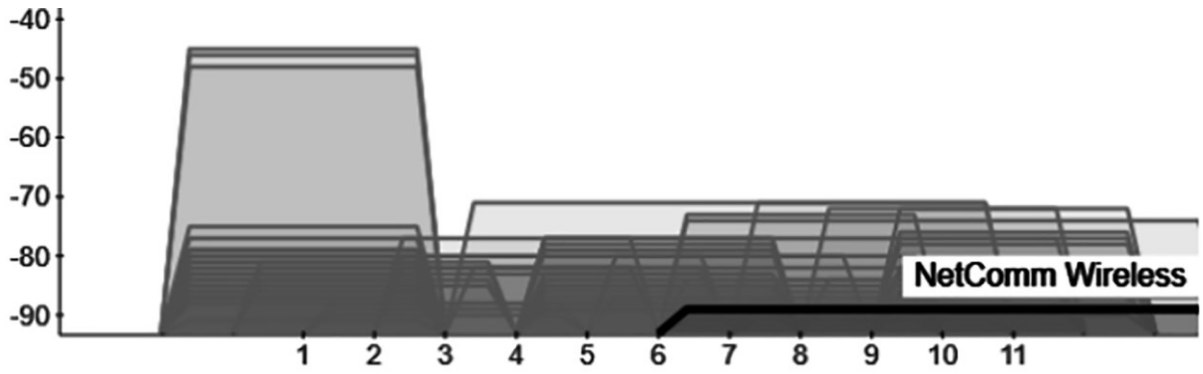


Figure 1. A typical example of, per AP, the signal strength (dBm) vs. Wi-Fi channel number in the 2.4 GHz band, as observed in a densely populated apartment block.

But figures such as Figure 1 do not evidence actual congestion. Wi-Fi systems apply a Media Access Control (MAC) communication protocol with which they "listen-before-talk", i.e. they try to send traffic only when they perceive the media as being "free". Congestion happens when an AP perceives all channels as being occupied with traffic virtually all the time. Wi-Fi's MAC protocol then still allows that AP to transmit, due to its random-back-off mechanism, but ultimately the total capacity will be shared with the other APs, and the performance (achievable throughput) of every individual AP will go down.

Congestion is not the same thing as interference. In the case of congestion, the AP recognises the received signal as Wi-Fi traffic emitted by other APs (or by devices communicating with those other APs). The MAC layer protocol then steps in and controls when APs and devices that can sense each other may have access to the media. When the AP does not recognise the received signal as Wi-Fi traffic, it registers it as interference or noise. Interference is typically caused by ISM devices in the neighbourhood, or devices using a different communication protocol such as Bluetooth or Zigbee (Zhang & Shin, 2011), or Wi-Fi devices emitting on different but overlapping channels.

An important property of commons is that if access to it is unlimited, it may lead to the so-called Tragedy of the Commons (Lloyd, 1833; Hardin, 1968). This is a situation within a shared-resource system where individual users acting independently according to their own self-interest behave contrary to the common good of all users by depleting or spoiling that resource through their collective action. As an apartment block is typically inhabited by a limited number of users, this situation does not seem to apply. However, the amount of devices that each user is operating at any given time is virtually unlimited, and if these devices interfere with each other, the Tragedy will take effect: all devices will operate at their maximum transmit power level in order to achieve an acceptable performance, and as such, ruin the media for each other.

http://doi.org/10.18080/ajtde.v5n2.93

Contrary to popular believe and what many authors have assumed so far (De Vries et al, 2013; Reed & Lansford, 2013; Nguyen et al, 2016), the effect called "Wi-Fi congestion" is *not* an example of the Tragedy. Adding APs to an apartment block does not *per se* reduce the overall Wi-Fi system capacity. As explained before, it rather just redistributes the available capacity over the operating APs and devices. In many cases it even enlarges the total available capacity slightly, as not every AP will be interfering equally as much everywhere in the apartment block. Certainly, if resident A has two APs, and resident B has only one, resident A may enjoy twice the capacity of resident B, causing resident B to buy his own additional AP. But that in itself does not deplete the commonly available resource.

But then, in 2013, Ozyagci, Sung and Zander (Ozyagci et al, 2013) showed that a system consisting of a continuously increasing number of Wi-Fi APs in an indoor environment will ultimately end up in an *over*-congested state. The difference between mere congestion and over-congestion is schematically depicted in Figure 2, which we copied from Ozyagci et al's article. A Wi-Fi system in an over-congested state uses an increasingly larger portion of the total available capacity for control traffic generated by the MAC protocol trying to mitigate the traffic congestion and packet collisions. The end result is actual depletion of the common resource. We therefore conclude that the Tragedy of the Commons does not apply to Wi-Fi congestion, but it does to Wi-Fi over-congestion.

Although Wi-Fi over-congestion is less common than Wi-Fi congestion, we recently showed that even if residents have only one AP and a few devices in operation per apartment, the system can be in a state of over-congestion at peak times, if the frequency planning is done badly (Den Hartog, Popescu et al, 2016). The conclusion for the short term is evident: operators and users should immediately stop trying to solve their Wi-Fi performance problems in densely populated areas by just adding APs and repeaters.

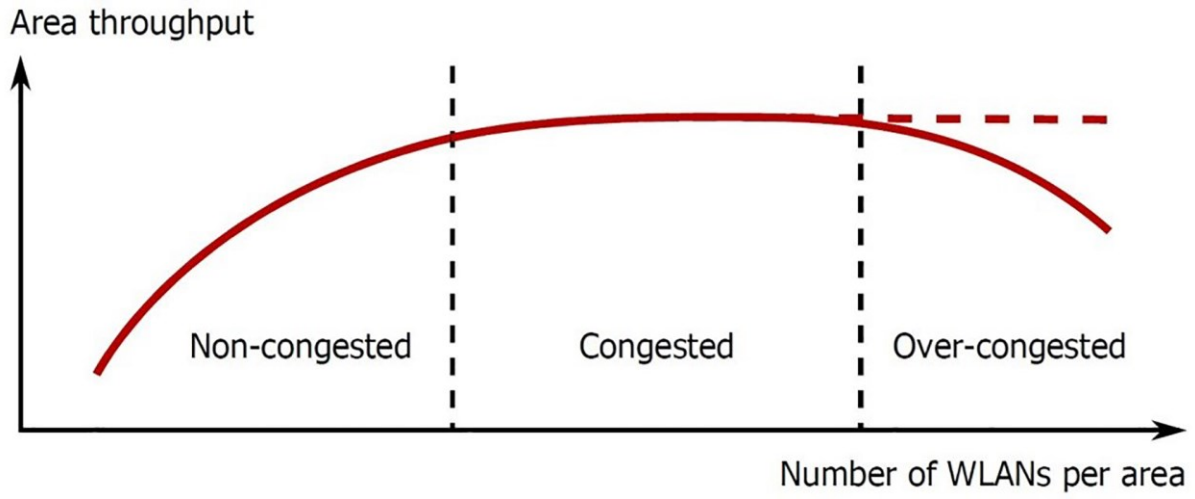


Figure 2. Wi-Fi over-congestion: decline of the area throughput (total available bits/s/m²) of an indoor Wi-Fi system consisting of a number of APs (WLANs) with increasing number of APs beyond the point of congestion (Ozyagci et al, 2013).

http://doi.org/10.18080/ajtde.v5n2.93

It is also a popular belief that migrating Wi-Fi systems to the 5 GHz frequency band and adding MIMO (Multiple-Input Multiple-Output) technology will alleviate the situation (Reed & Lansford, 2013). However, as long as vendors are not inclined to enable operation over the full frequency range available in the 5 GHz band, that alleviation will be short lived. In Figure 3 we show the observed signal strength vs. Wi-Fi channel number in the 5 GHz band, as observed in a high-rise building in The Hague, The Netherlands. Apparently, many channels are already occupied by multiple APs. Besides, as long as the demand in Wi-Fi connectivity seems to be growing faster than governments are adding capacity in the form of Class Licence spectrum, over-congestion will be a reality. Interestingly, the current approach of the industry to mitigate this issue is to improve frequency agility techniques (Sarijari et al, 2014), i.e. have a system automatically avoid frequencies which are already in use. This is not going to solve anything. It just surrenders the commons to the first and strongest players.

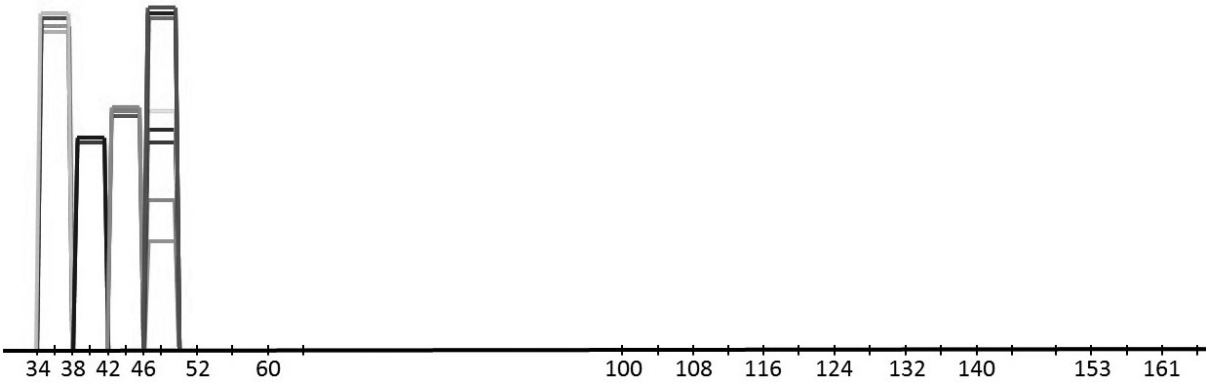


Figure 3. Relative signal strength vs. Wi-Fi channel number in the 5 GHz band, as observed in the head office of TNO in The Hague, The Netherlands.

Collaboration

A simple example: frequency selection

Most scholars studying commons economics agree that the first step towards a solution should be to stimulate players to design a form of self-regulation, i.e. come to a consensus about how the resource can be distributed as fairly as possible. How this could work in practice is illustrated by the following example. Imagine two close neighbours, residents A and B, both operating a Wi-Fi WLAN. Assume that the only parameter they have under their control is the selection of the frequency channel. Residents A and B observe that they both operate their WLAN at overlapping frequency channels. They decide to come to a consensus on who is going to reconfigure their AP to another, non-overlapping channel.

http://doi.org/10.18080/ajtde.v5n2.93

From a regulatory point of view, a number of questions then arise, including:

1) What is the legal status of such a deal; can it be enforced?
2) Imagine that residents A and B run competing businesses from home; is such a deal then legally allowed from an anti-trust perspective?
3) Imagine that resident A runs a business from home, and resident B does not; is resident A allowed to pay resident B to obtain an exclusive right to use a particular frequency, and how can the right price for such a deal be determined?

Asking these questions seems to be rather exaggerated in the context of this simple example, especially from an anti-trust perspective as regulators routinely target only large companies, but the reality of an apartment block is far more complicated.

Introducing the Spectrum Usage Broker and the Wi-5 System Operator

In the case of an apartment block, there may be more than a hundred residents, who do not necessarily know each other or have good relationships with each other. They may not be technically knowledgeable enough to be able to configure their AP. And apart from frequency channel, there are various other parameters that could be tuned in order to optimise the use of the spectrum, including transmit power and the ability to hand over devices to other, more suitable APs (horizontal handover) or even to a mobile network (vertical handover, reverse off-loading). Including these parameters as variables makes the process of consensus formation and the actual execution of the policy a task too complex to achieve without professional and automated assistance.

In the Wi-5 project we therefore introduced a generic business model with two new actors or business roles (Den Hartog, Kempker et al, 2016). They are the Spectrum Usage Broker and the Wi-5 System Operator. The Spectrum Usage Broker devises and maintains sensible spectrum sharing strategies between AP operators in a cooperative context. This may include a pricing agreement. The Wi-5 System Operator is in charge of operating a technology platform needed to automate the execution of the spectrum sharing strategies as devised by the Spectrum Usage Broker. An architecture of such platform is provided in the Wi-5 project deliverable D2.4 (Bouhafs, 2015).

In the use case of the dense apartment block, the new business roles could be implemented as follows. An apartment block has many other commons available to the tenants: hallway, joint garden, parking lot, etc. (depending on the details of the arrangement they may be club goods, but this does not alter our line of reasoning). Spectrum can be dealt with just as the other commons: tenants make mutual agreements about its use, and a caretaker has to execute the

http://doi.org/10.18080/ajtde.v5n2.93

agreements. The making of the mutual agreements can be facilitated by the Owners' Corporations or the building's Body Corporate, which is often an official entity, and tenants already pay a mandatory yearly subscription fee to their Corporation. The Corporation thus fulfils the role of Spectrum Usage Broker, and tries, within the bounds of regulation, to broker fair shares of the spectrum for every Local AP Manager. It may be aided in this task by a Code of Practice to be developed by, for instance, the Australian telecommunications self-regulatory body Communications Alliance.

After the Spectrum Usage Broker successfully matches the offer and demand of spectrum / capacity in the apartment block, the resulting policy is then handed to the Wi-5 System Operator, i.e. an entity that can control the individual APs. This could be an independent subcontractor, e.g. an IT company specialised in running a Wi-5-type platform, possibly "as a service" from the cloud. It could also be one of the broadband access providers servicing the apartments. Many access providers already have the knowledge and technology in place to take up such an additional role.

Regulation

Spectrum access

As said before, the use of, or access to, Class Licence spectrum for communication services is regulated in the national laws for telecommunications. In Australia, this is the Radiocommunications (Low Interference Potential Devices) Class Licence 2015 (ACMA, 2016). In the European Union, Commission Decision 2006/771/EC (EC, 2006) and Commission Recommendation 2003/203/EC (EC, 2003) apply. Two important aspects of the access rights are:

1. Usage on a non-protected basis: operators should accept the risk of interference between different users;
2. The use of the spectrum is not subject to individual rights; the spectrum is made available on a non-exclusive basis.

Stated differently, everybody has the fundamental right to access the spectrum anytime and anywhere, but has to accept possible interference from other users. This means that, in our use case of the apartment block, entrants such as new residents cannot be forced to participate in the collaboration, as this would equate to making the spectrum excludable, i.e. turning it into a club good where the Owners' Corporation is the club. This means that entrants should be enticed rather than forced to participate in the collaboration scheme. While assuming that all players act rationally, this will be achieved if joining the collaboration leads to lower operational costs and/or better network performance for all players involved. Said otherwise,

http://doi.org/10.18080/ajtde.v5n2.93

their business case should be attractive. The Wi-5 project is currently carrying out simulations applying a combination of cooperative and non-cooperative game theory to establish the benefits of defecting the collaboration, and to find out how many defectors our collaboration scheme can handle before it falls apart completely. Preliminary results indicate that collaborating is always beneficial in terms of obtained network performance (Van Heesch, 2016).

Another way of making the collaboration attractive is by monetising the right to access, i.e. resident A paying resident B to obtain an exclusive right to use a particular frequency. Although this may effectively lubricate the collaboration, resident A cannot force resident B to abstain from using resident A's frequency. This is because resident A cannot claim any individual rights, even though they paid for it. Said otherwise, it is not possible to turn the common good into a private good. In 2014, this has been confirmed by the FCC in the US, which did not allow Marriot International Incorporated to interfere with and disable Wi-Fi networks established by consumers in one of Marriot's conference centres, as Marriot claimed that within their premises consumers should only be allowed to connect to Marriot's Wi-Fi networks (FCC, 2014).

Anti-trust

Finally, we need to consider whether a collaboration as described is legally allowed, even though it cannot be enforced. From the point of view of antitrust regulation, this only concerns the cases where the roles of Spectrum Usage Broker or Wi-5 System Operator are taken up by large market players such as nationally operating service providers. In the use case of large apartment blocks, this can be realistically expected for the role of Wi-5 System Operator.

In the European Union, antitrust is defined in the Treaty of Lisbon, Article 101, (EC, 2008) which prohibits "agreements or concerted practices that limit or control production, or share sources of supply among undertakings". The role of the Spectrum Usage Broker is to assign a source of supply (spectrum) to the different Local AP Managers in an area, which suggests law infringement. However, clause 3 of article 101 makes an exception for such "agreements or concerted practices which contribute to improving the production or distribution of goods, while allowing consumers a fair share of the resulting benefit, albeit that such agreements or concerted practices do not limit or even eliminate competition in respect of the product."

This exception seems to be applicable to the case at hand: collaboration enlarges the size of the available Wi-Fi resources, allowing wireless users a fair share of the resulting benefit, and as such contributing to the improvement of the production of goods for which Internet access is needed, without limiting competition in respect of the product. Nobody is suffering from this collaboration. In Australia, anti-trust legislation is provided by the Competition and

http://doi.org/10.18080/ajtde.v5n2.93

Consumer Act 2010 (Cth) (CCA, 2010), and the cartel provisions in Part IV of the CCA amount to the same effect.

A final consideration is the risk of monopoly formation, as there is only one Spectrum Usage Broker and one Wi-5 System Operator in any given system. This risk can be mitigated by requiring these roles to be oversighted and/or executed by a not-for-profit organisation with transparent governance. In the case of the Owners' Corporation taking up the role of Spectrum Usage Broker, these requirements seem to come naturally. However, broadband access providers acting as Wi-5 System Operators may require such additional measures to be taken.

Conclusions

In this article we investigated the role of regulation in preventing Wi-Fi over-congestion in densely populated areas such as apartment blocks. From an analysis of the current trends in urbanisation, the number of Wi-Fi devices in use, the introduction of other technologies using the Class Licence bands, and the current approach that the industry takes to improve Wi-Fi system performance, we conclude that Wi-Fi over-congestion is unavoidable. Worse still, the currently popular strategy to solve performance problems by just adding APs and repeaters is only aggravating the problem in densely populated areas. Telecommunication service providers should immediately stop offering these "solutions" to their city-dwelling customers.

Over-congestion can only be avoided by having the relevant AP operators collaborating with each other. This follows directly from our conclusion that, in contrast to public Wi-Fi, Wi-Fi in apartment blocks is a true commons to which the Tragedy of the Commons applies. Here, we also make a clear distinction between congestion, over-congestion, and interference. Contrary to what is suggested in the literature, congestion is not depleting the resource. It just redistributes it.

The current regulations regarding spectrum access and anti-trust do not inhibit such collaboration, but they make it impossible to enforce it. Participation should be voluntarily, and AP operators should be enticed to collaborate by means of a positive business case. Preliminary results from the Wi-5 project indicate that this is feasible. We therefore conclude that, in contrast to earlier papers discussing Wi-Fi regulation (De Vries et al, 2013; Weiser & Hatfield, 2005), that further regulation is most likely not needed. However, the actors devising, maintaining, and executing sensible spectrum sharing strategies between AP operators have a monopoly position and should be oversighted or regulated in the cases where these roles are taken up by large market players. We propose that Codes of Practices are to be developed for these roles by, for instance, the Communications Alliance.

http://doi.org/10.18080/ajtde.v5n2.93

Acknowledgements

This work has been carried out in the framework of the Horizon 2020 Wi-5 project, which is partly funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement no. 644262.

References

ABS. 2016. 'Building Activity, Australia, Mar 2016'. 8752.0. Australian Bureau of Statistics.

ACMA. 2016. 'Radiocommunications (Low Interference Potential Devices) Class Licence 2015, Compilation No. 1'. Available from: https://www.legislation.gov.au/Details/F2016C00432 .

Anker, Peter. 2017. 'From spectrum management to spectrum governance'. *Telecommunications Policy*, (2017). DOI http://dx.doi.org/10.1016/j.telpol.2017.01.010 .

Bouhafs, Fayçal. 2015. 'Wi-5 interim architecture'. Deliverable 2.4 of the Horizon 2020 Wi-5 project. Available from http://www.wi5.eu/wp-content/uploads/2015/02/D2_4-Wi-5-Initial-Architecture-final-1.pdf .

Buchanan, James. 1965. 'An economic theory of clubs'. *Economica*, 32: pp. 1-14.

CCA. 2010. 'Competition and Consumer Act 2010 (Cth)'. Available from http://www.austlii.edu.au/cgi-bin/download.cgi/au/legis/cth/consol_act/caca2010265 .

Den Hartog, Frank; Kempker, Pia; Raschella, Alessandro; Seyedebrahimi, Mirghiasaldin. 2016. 'Network Uberization'. Available from https://www.slideshare.net/secret/JzIFRIPkLXS5Zz .

Den Hartog, Frank; Popescu, Alex; Djurica, Miodrag; Kempker, Pia; Raschella, Alessandro; Seyedebrahini, Mirghiasaldin; Arsal, Ali. 2016. 'Wi-Fi Optimisation Solutions Roadmap'. Deliverable 2.2 of the Horizon 2020 Wi-5 project. Available from http://www.wi5.eu/wp-content/uploads/2015/02/D2_2-Wi-Fi-Optimisation-Solutions-Roadmap.pdf .

De Vries, Jean Pierre; Simic, Ljiljana; Achtzehn, Andreas; Petrova, Marina; Mähönen, Petri. 2013. 'The emperor has no problem: Is wi-fi spectrum really congested?' In Proceedings of the 41st Research Conference on Communication, Information and Internet Policy (TPRC 41). 27-29 September 2013; Arlington, Virginia.

EC. 2003. 'Commission Recommendation of 20 March 2003 on the harmonisation of the provision of public R-LAN access to public electronic communications networks and services in the Community (Text with EEA relevance)' (2003/203/EC). *Official Journal of the European Union*, L 078: pp. 12 – 13. Available from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003H0203:EN:HTML .

EC. 2006. 'Commission Decision of 9 November 2006 on harmonisation of the radio spectrum for use by short-range devices' (2006/771/EC). *Official Journal of the European Union,* L 312: pp. 66-70. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:312:0066:0070:EN:PDF .

EC. 2008. 'Consolidated Versions of the Treaty on European Union and the Treaty of the Functioning of the European Union'. (2008/C 115/01). *Official Journal of the European Union*, C115/1. Available from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2008.115.01.0001.01.ENG&toc=OJ:C:2008:115:TOC#C_2008115EN.01004701 .

FCC. 2014. Order File No.: EB-IHD-13-00011303, Acct. No.: 201532080001, FRN: 0022507859, FRN: 0006183511, 3 October 2014. Available from https://www.fcc.gov/document/marriott-pay-600k-resolve-wifi-blocking-investigation .

Goggin, Gerard. 2014. 'New ideas for digital affordability. Is a paradigm shift possible?' *Australian Journal of Telecommunications and the Digital Economy*, Volume 2, Number 3, Article 42.

GPO. 2016. 'Title 47—Telecommunication', Code of Federal Regulations, pp. 845-964, US Government Publishing Office. Available from https://www.gpo.gov/fdsys/pkg/CFR-2016-title47-vol1/pdf/CFR-2016-title47-vol1-chapI.pdf .

Hardin, G. 1968. 'The tragedy of the commons'. *Science*, 162(3859): pp. 1243-1248.

IEEE. 2016. 'IEEE 802.11-2016. IEEE Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications', IEEE. Available from: http://www.techstreet.com/ieee/standards/ieee-802-11-2016?product_id=1867583 .

Lambert, Alex; McQuire, Scott; Papastergiardis, Nikos. 2014. 'Public Wi-Fi'. *Australian Journal of Telecommunications and the Digital Economy*, Volume 2, Number 3, Article 45.

Lloyd, W F. 1833. *Two lectures on the checks to population,* Oxford University, England.

http://doi.org/10.18080/ajtde.v5n2.93

Melbourne 2030. 2002. *Melbourne 2030; planning for sustainable growth*, Department of Infrastructure, State of Victoria. Available from https://www.planning.vic.gov.au/__data/assets/pdf_file/0018/20466/Melbourne-2030-Planning-for-sustainable-growth-text-only-version.pdf

Nguyen, T; Zhou, H; Berry, R A; Honig, M L; Vohra, R. 2016. 'The Cost of Free Spectrum', *Operations Research*, 64(6), 1217-1229.

Ozyagci, A; Sung, K W; Zander, J. 2013. 'Effect of propagation environment on area throughput of dense WLAN deployments'. In Proceedings of the Globecom Workshops (GC Wkshps). 9-13 December 2013; Atlanta, Georgia: 333-338.

Plummer, Daryl C; Reynolds, Martin; Golvin, Charles S; Young, Allie; Sullivan, Patrick J; Velosa, Alfonso; Lheureux, Benoit J; Frank, Andrew; Tay, Gavin; Bhat, Manjunath; Middleton, Peter; Unsworth, Joseph; Valdes, Ray; Furlonger, David; Goertz, Werner; Cribbs, Jeff; Beyer, Mark A; Linden, Alexander; Elkin, Noah; Heudecker, Nick; Austin, Tom; McIntyre, Angela; Chesini, Fabio; LeHong, Hung. 2016. '"Top strategic predictions for 2017 and beyond: Surviving the storm winds of digital disruption'. Gartner report G00315910. Gartner, Inc.

Potts, Jason. 2014. 'Economics of Public WiFi'. *Australian Journal of Telecommunications and the Digital Economy*, Volume 2, Number 1, March 2014, pp 1-9.

Reed, David P; Lansford, Jim. 2013. 'Wi-Fi as a Commercial Service: New Technology and Policy Implications' In Proceedings of the 41st Research Conference on Communication, Information and Internet Policy (TPRC 41). 27-29 September 2013; Arlington, Virginia.

Sarijari, M A; Abdullah, M S; Lo, A; Rashid, R A. 2014. 'Experimental studies of the ZigBee frequency agility mechanism in home area networks'. In Proceedings of the 2014 IEEE 39th Conference on Local Computer Networks Workshops (LCN Workshops), IEEE: pp. 711-717.

Speta, James B. 2008. 'Spectrum Policy Experiments: What's Next?' *University of Chicago Legal Forum*, (2008) Article 10. Available from: http://chicagounbound.uchicago.edu/uclf/vol2008/iss1/10

Telecommunications Act 1997. 2016. 'Telecommunications Act 1997 No. 47, 1997, Compilation No. 84'. Available from: https://www.legislation.gov.au/Details/C2016C00845 .

http://doi.org/10.18080/ajtde.v5n2.93

United Nations. 2015. 'World Urbanization Prospects: The 2014 Revision'. ST/ESA/SER.A/366. United Nations, Department of Economic and Social Affairs, Population Division.

Van Heesch, M P P. 2016. 'Combined Cooperation and Non-Cooperation for Channel Allocation and Transmission Power Control'. Available from: http://www.slideshare.net/MaranvanHeesch/thesis-van-heesch

Weiser, Philip J; Hatfield, Dale N. 2005. 'Policing the spectrum commons'. *Fordham Law Review* 74, (2005): pp 663-694.

Wi-5. 2015. Web page of the European Horizon 2020 Wi-5 project. At https://www.wi-5.eu .

Wi-Fi Alliance. 2017. 'Wi-Fi Alliance® publishes 7 for '17 Wi-Fi® predictions'. [Internet]. Wi-Fi Alliance Accessed 28 February 2017. Available from: http://www.wi-fi.org/news-events/newsroom/wi-fi-alliance-publishes-7-for-17-wi-fi-predictions .

Zhang, X; Shin, K G. 2011. 'Enabling coexistence of heterogeneous wireless systems: Case for ZigBee and WiFi'. In Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing, ACM.

http://doi.org/10.18080/ajtde.v5n2.93

Implementation of PCC-OFDM on a software-defined radio testbed

Gayathri Kongara
Monash University

Jean Armstrong
Monash University

Abstract: A software-defined radio implementation of polynomial cancellation coded orthogonal frequency division multiplexing (PCC-OFDM) on a field programmable gate array (FPGA) based hardware platform is presented in this paper. Previous publications on PCC-OFDM have demonstrated that, in comparison to normal cyclic prefix based OFDM, it is robust in the presence of many impairments including carrier frequency offset, multipath distortion and phase noise. The error performance of the two multicarrier techniques is compared on a practical wireless channel under common channel impairments such as carrier frequency offset, multipath and noise. Based on the comparative results obtained on the hardware platform, the properties of PCC-OFDM make it a suitable candidate for consideration in future 5G applications requiring robust performance in asynchronous environments with minimal out of band spectral emissions.

Keywords: PCC-OFDM, software-defined radio, 5G waveforms

Introduction

Many new applications such as the Internet of things (IoT) and machine-type communications are envisaged to be integral to fifth generation (5G) communication systems. These new scenarios in next generation systems will pose challenges that include an increase in aggregate date rate by 1000 fold, a requirement for round trip latency less than 1ms, and support for asynchronous communications (Aminjavaheri, Farhang, RezazadehReyhani, & Farhang-Boroujeny, 2015). Cyclic prefix orthogonal frequency division multiplexing (CP-OFDM) at present is the most successful physical layer solution in 4G communication systems. Due to frequency domain based signal processing, implementation of CP-OFDM results in significantly lower computational complexity for dispersive channels than single carrier transmission which is traditionally combined with time domain based receiver signal processing. However, CP-OFDM in its basic form will not meet the diverse requirements of

http://doi.org/10.18080/ajtde.v5n2.98

future 5G scenarios (Aminjavaheri et al., 2015; Kténas, 2015). One of the major limitations of CP OFDM is that without subcarrier (individual or a group) filtering it has high out-of-band (OOB) emissions (Aminjavaheri et al., 2015; Fettweis, Krondorf, & Bittner, 2009). CP-OFDM is also known to be highly sensitive to time and frequency synchronisation errors. The performance of CP-OFDM with practical time and frequency synchronisation techniques can deviate significantly from that obtained in ideal channel conditions. In addition, the length of the CP required in high delay spread channels increases as the number of samples affected by the inter-symbol-interference (ISI) increases. In general, the transmitter and receiver carrier frequencies are generated by different local oscillator clocks which may not have the same frequency accuracy and hence a non-zero carrier frequency offset (CFO) manifests in the signal reception. CP-OFDM based 4G systems rely on tight time and frequency synchronization to achieve the required error performance. However, this may not be realistic in future 5G systems. Inaccurate synchronization coupled with mobility in the wireless channel (Doppler) can disrupt the orthogonality of the sub-carriers in OFDM systems ultimately deteriorating the overall error performance. In some applications such as machine type communications (MTC) in 5G, maintaining strictly synchronized transmission is not possible as it is throughput inefficient. Tolerance to some asynchronous operation is very important for a good system design for such applications.

A diverse range of applications envisioned for 5G networks require access to spectrum within three key frequency ranges which are sub- 1 GHz, 1-6 GHz and above 6 GHz (This includes spectrum above 24 GHz). Compared to 4G frequency bands, 5G networks use very high frequencies and hence suffer from higher losses. Although OFDM is the dominant PHY layer solution in 4G applications, some of its weaknesses such as sensitivity to frequency offset errors and higher OOB emissions may become major issues for use with future 5G systems. Hence, there is a growing interest from both industry and academia in exploring alternative waveforms to suit 5G. A summary of OFDM waveform alternatives that rely on filtering of sub-carriers for 5G scenarios are currently under 5G networks (Farhang-Boroujeny & Moradi, 2016).

Polynomial cancellation coded orthogonal frequency division multiplexing (PCC-OFDM) modulation technique has been shown to have better spectral characteristics resulting in significantly lower OOB emissions than the conventional OFDM systems (Armstrong, 1999) (Armstrong, Gill, & Tellambura, 2000; Shentu, Panta, & Armstrong, 2003). The complex data samples in PCC-OFDM are mapped onto groups of subcarriers rather than individual subcarriers as in CP-OFDM. For example, a grouping order of two with PCC means that each data sample is mapped onto a pair of adjacent subcarriers scaled by factors $+1$ and -1. The overall OOB spectral roll-off for PCC-OFDM increases with increase in the grouping

http://doi.org/10.18080/ajtde.v5n2.98

order, however, this is achieved at the cost of some decrease in the spectral efficiency (Armstrong, 1999). However, some other properties of PCC-OFDM can reduce the overall spectral loss. The CP required in conventional OFDM can be a significant overhead in high delay spread channels unlike conventional OFDM, does not need a CP.

Another way to increase the spectral efficiency of PCC-OFDM is to overlap symbols before data transmission. With this approach, ISI is deliberately introduced into the signal before transmission. A multi-stage equaliser at the receiver which operates as a linear, or decision feedback equaliser is then used to recover the signal (Armstrong et al., 2000).

A number of alternative schemes are currently under development. These mainly rely on filtering the subcarriers with special pulse shaping filters; it was shown that the required low OOB emissions for 5G could be achieved with the use of long filters (Aminjavaheri et al., 2015; Farhang-Boroujeny & Moradi, 2016; Kténas, 2015). Filtering of individual sub-carriers or a group of sub-carriers decreases signal to noise ratio (SNR) and adds to the end-to-end latency of the system. Other techniques that implement sophisticated interference cancellation at the receiver along with CFO compensation have been explored (Defeng & Letaief, 2005; Lee & Lee, 2011). Such techniques again add signal-processing complexity and latency. Filtering based multicarrier waveforms of (Aminjavaheri et al., 2015; Fettweis et al., 2009) were shown to provide significant improvement compared to conventional CP-OFDM systems.

In this paper, we describe a system implementation of PCC-OFDM on the USRP hardware. Since the introduction of PCC-OFDM in 1998-99 (Armstrong, 1999), a number of studies describing the theoretical and numerical performance limits under various propagation impairments have appeared in the literature. However, there are no publications on PCC with hardware implementation results. In this paper, we present a description of PCC-OFDM and CP-OFDM implementations on a software-defined radio (SDR) test platform. We introduce time and frequency errors in the experiments to compare their relative strengths and weaknesses.

The organisation of the paper is as follows. In Section 2, the software defined radio platform is described. Section 3 introduces the PCC-OFDM system model. Section 4 describes the receiver processing algorithm for the time and the frequency synchronisation. Channel estimation and equalisation are discussed in Section 4. Section 5 presents performance results from the hardware implementation of PCC and CP-OFDM under various propagation scenarios.

http://doi.org/10.18080/ajtde.v5n2.98

Software Defined Radio Platform

The SDR platform consists of one or more USRP 2943R devices that can be programmed as a transmitter or receiver. The USRPs connect to a personal computer (PC) running National Instruments (NI) LabVIEW 2016 as shown in the two configurations in Figure 1. The high-speed low latency PCIeX4 interface cards installed on the PC are capable of data transfer at 800 Mbps (NI, 2015). The PCIeX4 interface card loads compiled bit files from the LabVIEW environment on to the FPGA motherboard. This host-based LabVIEW system implementation offers great flexibility in the system design, configuration and testing. In addition to the entire physical layer signal processing in LabVIEW, RF system parameters can be configured in software. The transmitter (TX) is connected to the USRP RIO0 device's RF0 daughter board. The receiver (RX) is on either a RF1 daughter board of the same USRP RIO0 as shown in Figure 1 (a) or a physically separate device named USRP RIO1 that is connected to the same PC but placed at 1m distance away from the USRP RIO0 as shown in Figure 1 (b).

The frequency accuracy is usually specified in the hardware specification sheets. This is calculated from the frequency error normalised to the sampling rate. For USRP RIO 2943R, a ± 2.5 parts per million (ppm) frequency error can occur (NI, 2015). This means that the hardware introduces ± 2.5Hz frequency error for a 1MHz sampling rate. This is much smaller than the subcarrier spacing we use in our experiments. The data sheets of USRP-RIO indicate that the hardware introduces a delay of less than 1 μsec. Hence, hardware induced impairments are so small when experiments are conducted on the same device that frequency synchronisation is typically not necessary.

However, in experiments where the transmitter and receiver are separated by a distance 1m or greater, a significant time and frequency offsets affect the received signal. A correlation based two-stage time and frequency synchronisation algorithm is described in this paper. Performance results for PCC and CP-OFDM implementations with the synchronisation algorithm are compared for frequency offset in an AWGN channel and various modulation orders

PCC-OFDM System Model

The transmitter block diagram of the PCC-OFDM is illustrated in Figure 2. At the transmitter, information bits are modulated using 16-QAM and 64-QAM with PCC carrier grouping. A grouping order of two is assumed in this paper, so QAM modulated samples are mapped onto a weighted pair of adjacent subcarriers with a relative weighting of $+1, -1$.

http://doi.org/10.18080/ajtde.v5n2.98

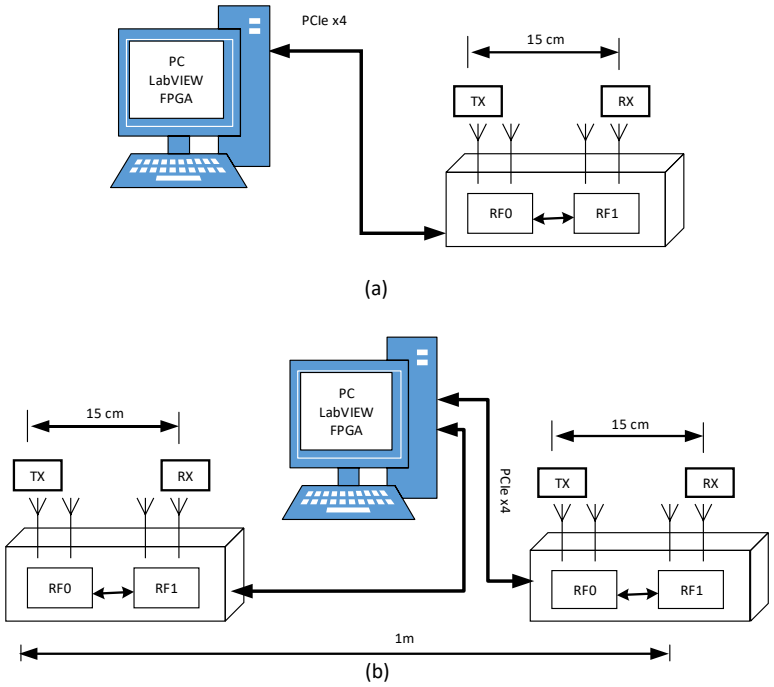


Figure 1: LabVIEW based Software Defined Radio Platform with (a) one USRP (b) two USRPs

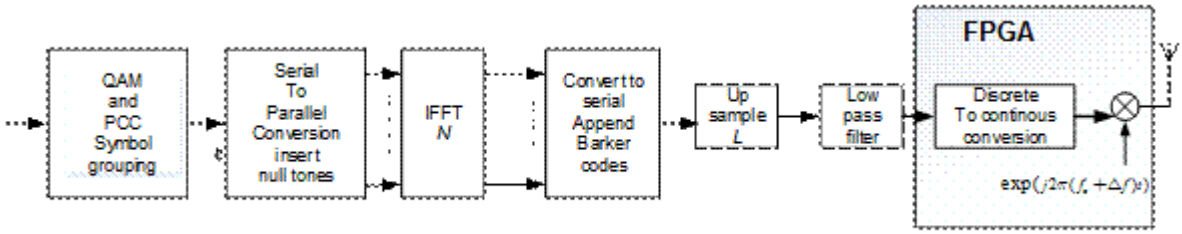


Figure 2: PCC OFDM Transmitter

A typical transmission frame carrying multiple OFDM or PCC symbols with Barker training sequences is shown in Figure 3. A repetition of training information induces periodicity into the transmission which the receiver uses in the estimation of CFO, delay and channel state information. Each training block is a Barker code which comprises a sequence of $+1$ or -1 and exhibits good aperiodic autocorrelation properties (Moose, 1994).

http://doi.org/10.18080/ajtde.v5n2.98

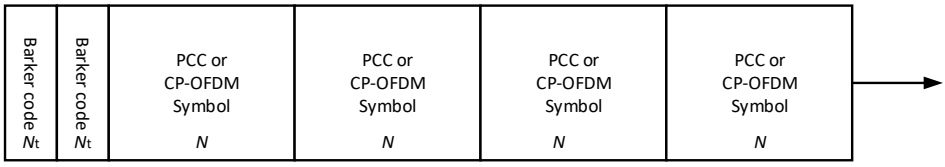


Figure 3 Transmission frame carrying PCC-OFDM symbols

The serial information QAM symbol stream is split into *N* parallel sub-streams that are mapped onto *N* subcarriers. The pair of subcarriers around Nyquist frequencies are nulled. Four subcarriers around zero frequency are further nulled as any DC bias complicates the analogue to digital and digital to analogue conversion processes. After the subcarrier nulling stage, the time domain data is generated using an inverse fast Fourier transform (IFFT). The parallel information streams are converted back into a serial data stream and up-sampled. This signal then passes through a band-limiting filter and then the RF front-end for digital-to analogue conversion and transmission across the wireless channel.

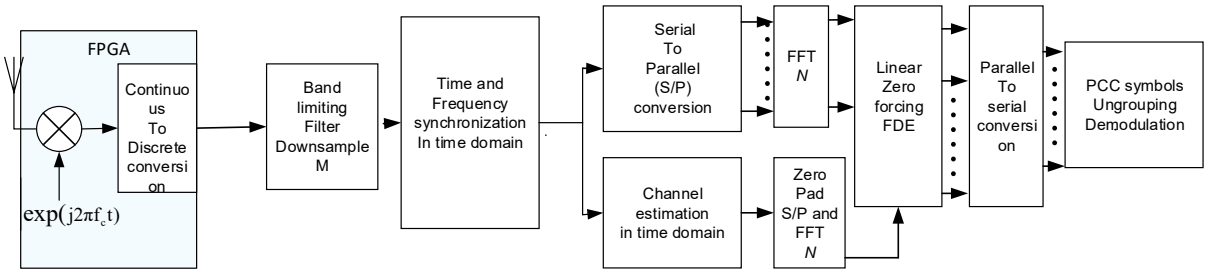


Figure 4 PCC OFDM Receiver

The block diagram of the implementation of the PCC-OFDM receiver is shown in Figure. 4. The time offset is denoted as *D*, multipath dispersion of the channel is defined by the vector $\mathbf{h} = \left[h(0),\ h(1),...h(p\ -\ 1)\right]$ and CFO by Δf. The CFO normalised to the sample period T is denoted by $\varepsilon = \Delta fT$. Given the transmit signal $x(n)$ (for the n^{th} discrete time instant) at the IFFT output, the received signal sample $y(n)$ is given by

$$y(n) = \exp(j2\pi\varepsilon n)\sum_{l=0}^{p-1} h(l)x(n-l-D) + \omega(n). \quad (1)$$

Where, $\omega(n)$ corresponds to the sample of AWGN and p is the multipath (wireless) channel order.

http://doi.org/10.18080/ajtde.v5n2.98

Time and frequency synchronisation

The delay parameter may span several symbol periods and must be estimated to determine the frame start. Prior to this, a fractional time synchronisation is carried out on the oversampled received frame. To achieve this the received OFDM or PCC signals are up-sampled with the same oversampling rate as was used at the transmitter. The calculation of the fine time synchronisation offset involves tracking the energy of the received frame over L fractional time delays where each delay is Ts/L. The time index corresponding to the maximum energy of the received frame is used to compensate the fractional time offset. The fine synchronisation results in a sample level signal that has an integer time offset. This is carried out at a sample level following fine synchronisation process. A simple correlation-based peak-detection technique is used to estimate the unknown integer delay. To estimate the integer time offset, the received signal corresponding to the training phase is correlated with the training signal given by

$$\mathrm{R}(n) = \left\| \sum_{k=0}^{N_t-1} t^*(k)\, y(n+k) \right\|^2, \tag{2}$$

where N_t is the length of the training sequence. An estimate of the delay parameter D in (1) is then calculated by taking the index n of the maximum value of the correlation metric given in (2). This is given by

$$\hat{D} = arg\left(\max_{\mathrm{n}} \mathrm{R}(n) \right) \tag{3}$$

The frame start is set to the estimated integer delay parameter $\hat{D}$ to compensate for the overall delay in transmission. However, in the presence of CFO, the correlation peak may be erroneously shifted due to ICI. For small estimation errors, a simple one-tap linear equaliser can be used to compensate for the offset. However, in the presence of significant CFO, the frequency error should be properly estimated and compensated for before carrying out further receive processing. After frame synchronisation using $\hat{D}$ the received signal in (1) reduces to

$$y(n) = \exp(j2\pi\varepsilon n) \sum_{l=0}^{p-1} h(l)x(n-l) + \omega(n), \tag{4}$$

The next step is to estimate the frequency error $\varepsilon = \Delta fT$. Noticing that the two received blocks corresponding to the consecutively transmitted training blocks each of length N_t differ by a phase of $\exp(j2\pi N_t\varepsilon)$, we formulate a least-squares problem. For brevity, we denote $\exp(j2\pi N_t\varepsilon) = a$ and the objective function is written as

http://doi.org/10.18080/ajtde.v5n2.98

$$J(a) = \sum_{n=0}^{Nt-1} \| y(n + N_t) - a y(n) \|^2 . \quad (5)$$

Minimisation of the metric defined in (5) results in

$$\hat{a} = \frac{\sum_{n=p}^{Nt-1} y(n + N_t)\, y^*(n)}{\sum_{n=p}^{Nt-1} \| y(n + N_t) \|^2} \quad (6)$$

In (6), the first few training samples affected by channel memory are discarded when calculating the synchronisation parameters. The normalised CFO estimate is simply the angle of (6) which is given by:

$$\hat{\varepsilon} = \angle \left(\frac{\sum_{n=p}^{Nt-1} y(n + N_t)\, y^*(n)}{2\pi N_t} \right) \quad (7)$$

The estimated $\hat{a}$ is used in compensating the CFO at the receiver, by multiplying by the received frame with $\hat{a} = \exp(-2\pi N_t \hat{\varepsilon})$. Synchronisation is followed by channel estimation and equalisation based on the training vector $t = \left[t(0),\ t(1),\ \ldots,\ t(N_t - 1) \right]$. The time domain signal after time and frequency synchronisation is given by

$$y(n) = \sum_{l=0}^{p-1} h(l)\, t(n-l) + w(n) \qquad n = [0,\ 1,\ \ldots, N_t - 1] \quad (8)$$

The channel effect in (8) should first be estimated. A simple and practical channel estimation algorithm based on the linear least squares (LS) criterion is employed. Unlike minimum mean squared error (MMSE) estimation approaches, LS does not require accurate real-time estimates of noise at the receiver. Channel estimates are calculated upon minimising the error between the received samples and the training samples ignoring the effect of additive noise. Hence, this approach has a disadvantage in high noise conditions where an implementation of MMSE channel estimates gives inaccurate estimates. The estimated channel coefficients in the time domain give the time domain equaliser parameters. The frequency domain equaliser (FDE) is calculated by Fourier transformation of the time domain channel parameters. A direct time domain approach to calculation of equaliser parameters is obtained by forming an error metric from the training and the estimated samples. The equaliser vector is calculated by

http://doi.org/10.18080/ajtde.v5n2.98

$$\mathbf{f} = (\bar{\mathbf{y}}^{*}\ \bar{\mathbf{y}})^{-1}\ \bar{\mathbf{y}}\mathbf{t} \tag{9}$$

where $\bar{\mathbf{y}}$ is a matrix of dimension $(N_t + 1)x\ (L + 1)$ and $\mathbf{f}$ is the equaliser vector with L taps of dimension $(L + 1)x1$. The discrete Fourier transform of (9) gives the FDE $\boldsymbol{F} = [\,F(0), F(1), \ldots, F(N\ -\ 1)]$. A simple zero-forcing FDE is implemented to compensate for the distortion due to multipath which is given by

$$X(k)\ = \frac{Y(k)}{F(k)} \qquad k\ =\ [0,\ 1,\ \ldots, N]. \tag{10}$$

After FDE, for PCC, the grouped subcarriers are de-multiplexed and decoded for data recovery and BER is calculation. In the following section hardware results are presented for the receiver described in this section.

Hardware Results

In this paper, we extend the experiments reported in our previous papers (Kongara & Armstrong, 2016, 2017) in which a performance comparison of the hardware implementations of PCC-OFDM and CP-OFDM with $4-$QAM was presented. To investigate the effects of CFO and multipath, we consider two experimental set-ups as illustrated in Figure 1. In Figure 1 (a) the physical separation of transmitter and receiver is 15 cm as they are connected to the same USRP. The set-up in Figure 1 (b) has transmitter and receiver on different devices but the range is restricted to 1m as both devices are connected to the same PC. The focus of our current work is comparing the robustness of the two modulation schemes in real wireless channels. Increasing the range to more than a meter is possible but is outside the scope of our current investigation.
In the current paper, experimental results for the hardware configuration shown in Figure 1 for 16-QAM and a 64-QAM formats are discussed. In all our experiments, a carrier frequency in the Industrial Scientific and Medical (ISM) band (2.41 GHz) and an antenna gain of 10 are used. A traditional time domain approach to synchronisation and channel estimation (see Section 4 above) is employed and hardware performance of the two systems is compared. The need for more accurate time synchronisation for 64-QAM compared to 16-QAM and 4-QAM schemes is discussed. We then compare the BER performance of the time and frequency synchronised PCC and OFDM systems under variable AWGN conditions, CFO, FFT sizes and transmission frame lengths. Both instantaneous and average results are used in understanding and drawing conclusions on the robustness of PCC system against normal OFDM system.

Experiments are conducted on a SDR platform with the transmitter and receiver separated by a distance of 15 cm (shown in Figure 1 (a)) and 1 metre (shown in Figure

http://doi.org/10.18080/ajtde.v5n2.98

1 (*b*)). We first investigate the requirement for oversampling to enable initial timing estimation by tracking the output energy of the detector described in Section 4. Each transmission frame is approximately 4 ms in duration and carries 12k and 24k information bits for the PCC-OFDM and CP-OFDM schemes, respectively. For 64-QAM and N=1024, the transmission frame carries 4 symbols modulated using either PCC or normal CP-OFDM symbols.

In Figure 5 (*a*) we see an example of a transmission frame containing the time domain PCC-OFDM symbols after IFFT operation that is generated with an FFT size of N=1024. The received constellation corresponding to the transmission of Figure 5 (*a*) is shown in Figure 5(*b*). Figure 5 (*c*) and (*d*) illustrate the OFDM transmit waveform and received constellation. A subcarrier grouping order of 2 is used for PCC-OFDM symbols in this paper. As can be seen in Figure 5 (*a*), due to the PCC coding, the symbol transitions are smooth with an effect of time domain windowing. This time domain symbol shape achieved with PCC grouping is shown to be equivalent to filtering with a Hanning window (Panta & Armstrong, 2003). The OFDM waveform without additional filtering has sharp transitions compared to PCC-OFDM. This results in higher out of band emissions for OFDM compared to that for PCC system. The spectral efficiency of PCC is halved due to the sub-carrier grouping. However, a CP length equal to a quarter of the symbol period is generally needed for OFDM to eliminate ISI. Whereas PCC is more tolerant to multipath propagation

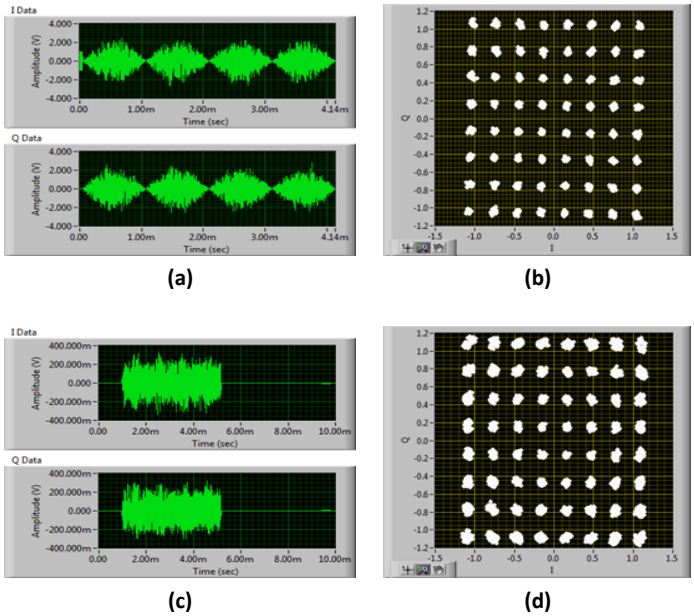
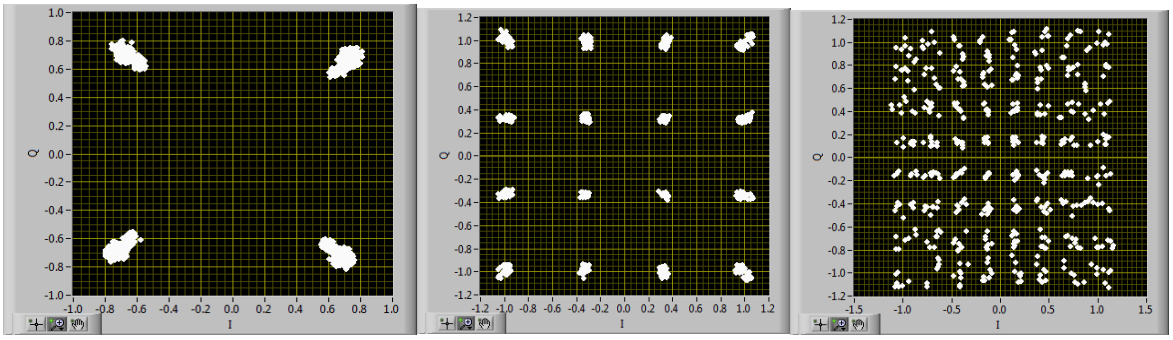


Figure 5: (a) I and Q components of the transmission frame carrying four PCC-OFDM symbols of 64QAM samples. (b) Received constellation corresponding to (a) over the wireless channel. (c) I and Q components of

http://doi.org/10.18080/ajtde.v5n2.98

a transmission frame carrying four CP-OFDM symbols of 64QAM samples. (d) Received constellation corresponding to (c) over the wireless channel.

In Figure 6, examples of the received constellations that correspond to the transmission of PCC-OFDM symbols carrying QPSK, 16-QAM and 64-QAM data are shown for an oversampling factor of 2. The average BER calculated from over 100 transmission frames result in a BER= 0 for 4-QAM . However, an average BER = 0.006185 for 16 QAM and BER =0.022 for 64-QAM is obtained on the real wireless channel with clear line-of-sight. As seen from the constellations, higher order modulations such as 16-QAM and 64-QAM demand more precise estimates for the timing information at the receiver.



(a) QPSK samples **(b) 16-QAM samples** **(c) 16-QAM samples**

Figure 6: Received constellation with an oversampling factor of 2 corresponding to the transmission of PCC frame carrying (a) QPSK samples (b) 16-QAM samples (c) 64-QAM samples.

In our hardware implementation, we used an oversampling approach similar to (Awoseyila, Kasparis, & Evans, 2009) to calculate more precise sampling instants. This approach is based on tracking the energy of the received frame over a certain range of delays. A sliding window is used that outputs the energy of the received frame for every sample delay. The tracking process continues until it finds the sample delay that corresponds to the maximum output energy. In Figure 7, the output energy for oversampling factors of L=M=2, 4 and 10 are illustrated. As can be seen from the plots, the constellations with higher oversampling factors result in more precise sample time offset. However, it should be noted that choosing higher oversampling factors consumes more hardware resources and a trade-off between the complexity and required accuracy for a system should be carefully considered.

http://doi.org/10.18080/ajtde.v5n2.98

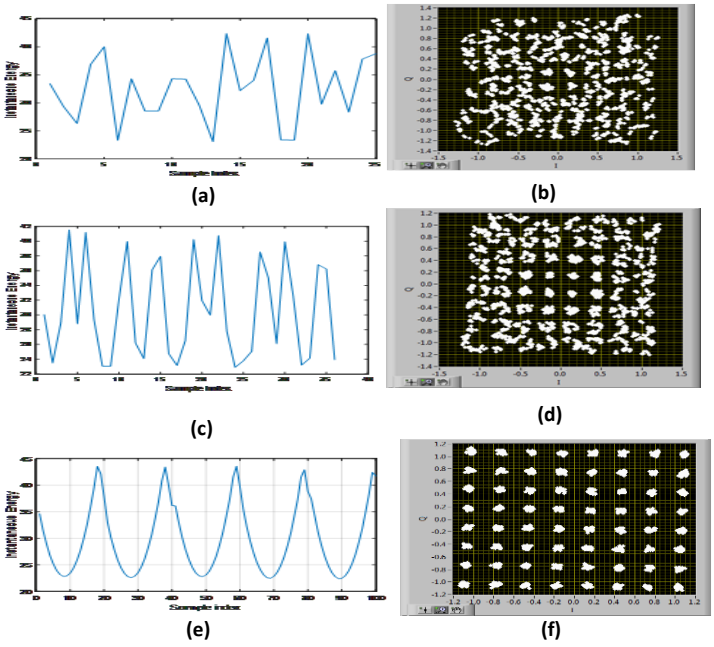


Figure 7: Output of the sliding energy detector with oversampling factor (a) L=M=2 (b) L=M=4 (c) L=M=10

In Figure 8, BER averaged over 100 channel realisations of transmission is plotted for a varying SNR. The experiment is over a real wireless channel with the transmitter and the receiver connected as in Figure 1 (a) to the same USRP device and sharing a common clock (no CFO). The channel is clear line of sight, and hence the channel estimator is programmed to estimate just one coefficient to compensate for fading in the line of sight path. Zero-forcing based FDE is calculated from the estimated channel and the signal is equalised in the frequency domain. The BER performance of PCC OFDM system is significantly better than the normal OFDM with a 4dB SNR improvement with good time synchronisation achieved with an oversampling factor of $L = M = 10$. However, a performance degradation is seen due to inaccurate timing synchronisation when $L = M = 2$ and 4 are used.

OFDM transmission through a fading wireless channel can disrupt the orthogonality of sub-carriers generating ICI and ISI. This makes the OFDM signal reception sensitive to frequency synchronisation errors. The CFO is estimated as described in Section 4 using the received transmission frames of a PCC or a normal OFDM.

Another important factor that can affect the average BER is the FFT size. In Figure 9(a) and (b) transmission frames with $N = 64$, and $N = 1024$ carrying PCC symbols is shown. The average BER with $N = 64$, and $N = 1024$ for PCC and OFDM is illustrated. With no added AWGN, the PCC scheme exhibits lowest average BER for $N = 1024$. Comparing PCC N=1024 with OFDM for a CFO range of 100Hz to 400Hz, OFDM performance is poorer than the former. A similar observation can be made from the performance obtained for $N = 64$. The

http://doi.org/10.18080/ajtde.v5n2.98

performance degradation for a smaller FFT size with $N = 64$ is more for the OFDM system compared to PCC. There is some loss in PCC performance with smaller FFT size of $N = 64$ compared to $N = 1024$. However, OFDM with $N = 1024$ is slightly worse than PCC with $N = 64$. This means that the CFO estimation error degrades OFDM more than PCC.

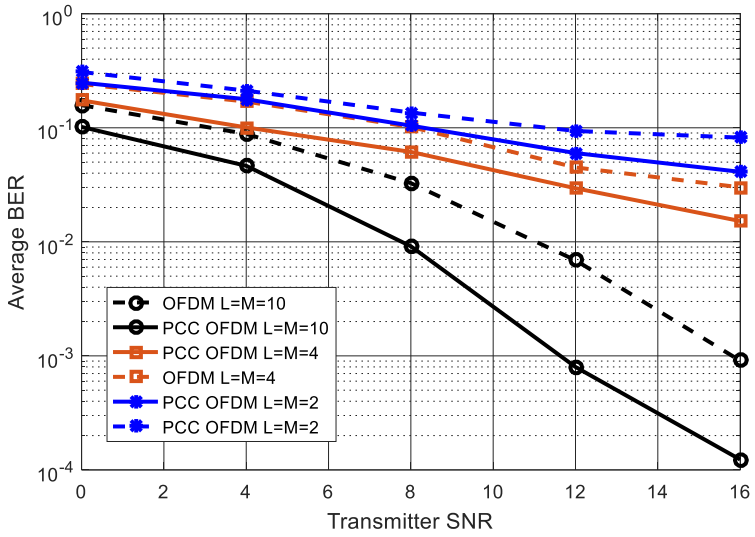
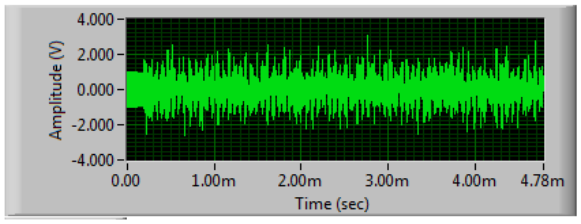


Figure 8 Transmit SNR versus BER comparison of PCC-OFDM against CP-OFDM for varying oversampling values.

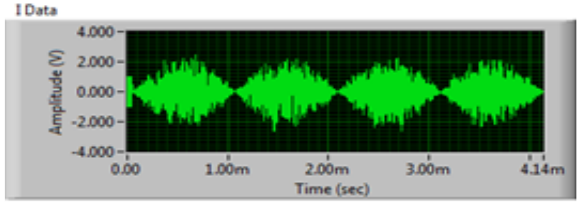
As noise can further degrade performance, the performance at various SNR and CFO are studied in Figure 10. One observation from the results shown in Figure 10 is that for SNR =15, 20 and no noise conditions and under different values of CFO scenarios, PCC significantly outperforms normal OFDM. However, for higher values of CFO, both OFDM and PCC result degraded error performance.

In the next experiment, we look at the effect of long transmission frames on the achieved BER performance. The separation distance between the transmitter and receiver is set to 1m and the transmitter and receiver are on two separate USRP devices. The experimental set-up is as shown in Figure 1 (b). Transmission frame time with the test set-up of Figure 1(b) is $1/6^{th}$ of the frame duration that of Figure 1 (a). This is because, for increased path lengths, the channel state information changes significantly over the duration of the transmission frame and the synchronisation and channel parameters estimated at the start of the frame can be outdated for processing long frames.

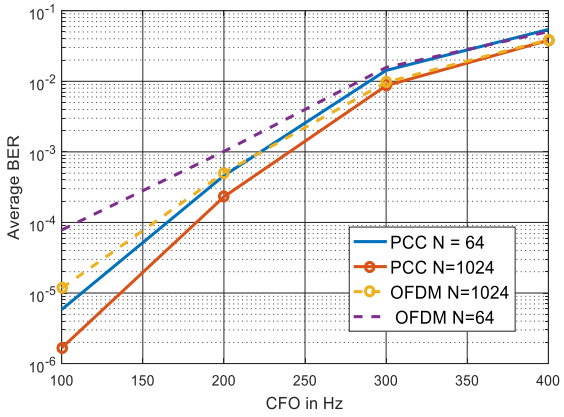
http://doi.org/10.18080/ajtde.v5n2.98



(a)



(b)



(c)

Figure 9: A transmission frame with PCC-OFDM 64 QAM samples (a) FFT size $N = 64$ (b) FFT size *N=1024* (c) Average BER vs CFO in Hz for PCC and OFDM of (a) and (b).

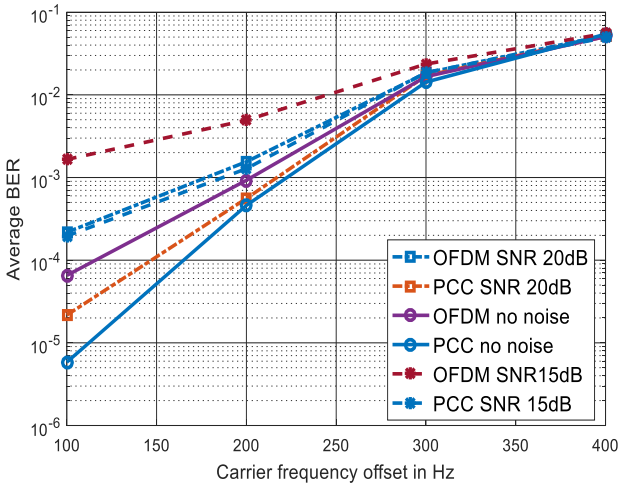


Figure 10. Average BER for comparison of PCC-OFDM against CP-OFDM for block FFT size *N=64* and SNR = 15 and 20. BER performance of the hardware with no additional noise is also shown for comparison.

http://doi.org/10.18080/ajtde.v5n2.98

In Figure 11, instantaneous observations of the received frames after synchonisation and FDE for frames carrying 16-QAM and 64-QAM samples are plotted for smaller frame duration of 660μsec. In Figure 12, the average BER calculated from 100 transmission frames of PCC and OFDM with varying number of symbols carried in each frame is shown. As expected, the BER performance of PCC is an order of a magnitude better than the normal OFDM system. As the transmission frame duration increases, BER performance of both systems degrade. However, PCC outperforms OFDM for all frame durations considered in hardware experiments. The correlation based time synchronisation algorithm discussed in Section 4 compensates for the delay introduced by the wireless channel but the performance degrades with increase in frame duration (indicated by increase in number of symbols). As the frame size increases, indoor wireless propagation induces more ISI and more frequent receiver parameter estimation is required to compensate the detrimental effects due to ICI and ISI.

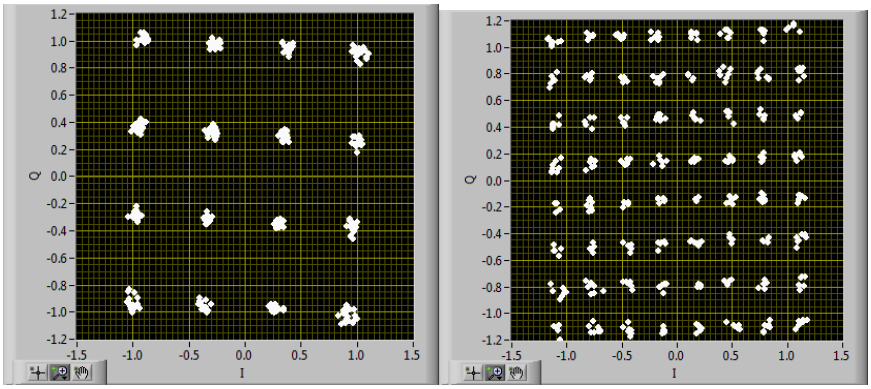


Figure 11 Instantaneous observations of transmission over 1m wireless channel after CFO compensation and equalisation

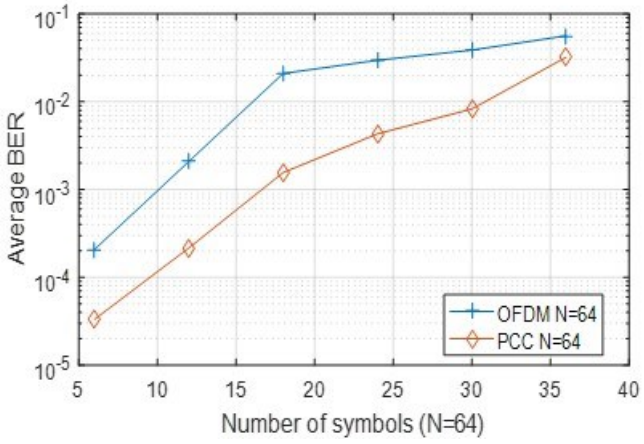


Figure 12. Average BER comparisons of PCC and OFDM symbols carrying 16 QAM samples over 1m wireless channel.

http://doi.org/10.18080/ajtde.v5n2.98

Conclusions

A hardware implementation of PCC-OFDM and CP-OFDM with symbols with 16-QAM and 64-QAM samples is presented in this paper. BER results from the hardware implementation of the two multicarrier approaches on USRP based SDR platform under variable channel impairments are compared. Hardware experimental results demonstrate the robustness of PCC-OFDM to time and frequency estimation errors under various noise conditions. A simple time domain correlation based timing and CFO estimation algorithm is implemented in LabVIEW to synchronise both systems. The hardware received time domain waveform of PCC-OFDM has smoother transitions resulting in lower OOB emissions than CP-OFDM. This is a major requirement for 5G systems. The performance degradation due to increased noise level for PCC is 3 – 4dB less than normal OFDM. The effect of smaller FFT size affects PCC less than OFDM. Experiments conducted on a wireless channel with a separation distance of 1 m show that the wireless channel introduces ICI and ISI in the received signals of PCC and OFDM. After synchronisation and equalisation, BER of PCC and OFDM compared under variable frame duration. PCC outperforms OFDM in for all frame sizes.

Acknowledgements

The authors wish to thank engineers at National Instruments for providing technical support on the SDR platform and Dr. Michael Biggar for sharing helpful comments and suggestions.

References

Aminjavaheri, A; Farhang, A; RezazadehReyhani, A; Farhang-Boroujeny, B. (2015). "Impact of timing and frequency offsets on multicarrier waveform candidates for 5G". Paper presented at the 2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE), 9-12 Aug. 2015.

Armstrong, J. (1999). "Analysis of new and existing methods of reducing intercarrier interference due to carrier frequency offset in OFDM". *IEEE Transactions on Communications*, 47(3), 365-369. doi: 10.1109/26.752816

Armstrong, J; Gill, T; Tellambura, C. (2000). "Performance of PCC-OFDM with overlapping symbol periods in a multipath channel". Paper presented at the Global Telecommunications Conference, 2000. GLOBECOM '00. IEEE.

Awoseyila, A. B; Kasparis, C; Evans, B. G. (2009). "Robust time-domain timing and frequency synchronization for OFDM systems". *IEEE Transactions on Consumer Electronics,* 55(2), 391-399. doi: 10.1109/TCE. 2009.5174399

Defeng, H; Letaief, K. B. (2005). "An interference-cancellation scheme for carrier frequency offsets correction in OFDMA systems". *IEEE Transactions on Communications,* 53(7), 1155-1165. doi: 10.1109/TCOMM.2005.851558

Farhang-Boroujeny, B; Moradi, H. (2016). "OFDM Inspired Waveforms for 5G".

IEEE Communications Surveys & Tutorials, 18(4), 2474-2492. doi: 10.1109/COMST.2016.2565566

Fettweis, G; Krondorf, M; Bittner, S. (2009). "GFDM - Generalized Frequency Division Multiplexing". Paper presented at the VTC Spring 2009 - IEEE 69th Vehicular Technology Conference 26-29 April 2009.

Kedia, M. (2005). "A computationally efficient method for estimating the channel impulse response for the IEEE 802.11b (WLAN)". Paper presented at the PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, 24-26 Aug. 2005.

Kongara, G; Armstrong, J. (2016). "Implementation of PCC OFDM on a software defined radio platform". Paper presented at the 2016 26th International Telecommunication Networks and Applications Conference (ITNAC), 7-9 Dec. 2016.

Kongara, G; Armstrong, J. (2017). "Performance evaluation of PCC OFDM on a software defined radio platform". Paper presented at the 2017 International Conference on Computing, Networking and Communications (ICNC), 26-29 Jan. 2017.

Kténas, D. (2015). 5G NOW: Final Assessment of Demonstrator Concept and Implementation

Lee, K; Lee, I. (2011). "CFO Compensation for Uplink OFDMA Systems with Conjugated Gradient". Paper presented at the 2011 IEEE International Conference on Communications (ICC) 5-9 June 2011.

Moose, P. H. (1994). "A technique for orthogonal frequency division multiplexing frequency offset correction". *IEEE Transactions on Communications, 42*(10), 2908-2914. doi: 10.1109/26.328961

National Instruments. (2015). Overview of the NI USRP RIO Software Defined Radio.

Panta, K; Armstrong, J. (2003). "Spectral analysis of OFDM signals and its improvement by polynomial cancellation coding". *IEEE Transactions on Consumer Electronics, 49*(4), 939-943. doi: 10.1109/TCE.2003.1261178

Shentu, J; Panta, K; Armstrong, J. (2003). "Effects of phase noise on performance of OFDM systems using an ICI cancellation scheme". *IEEE Transactions on Broadcasting, 49*(2), 221-224. doi: 10.1109/TBC.2003.810074

http://doi.org/10.18080/ajtde.v5n2.98

Interference to Telephone Lines

Simon Moorhead
Ericsson Australia & New Zealand

Abstract: A paper from 1936 exploring the effects of electrification of country Tasmania and the increasing interference to telecommunication circuits by high voltage power lines installed in close proximity.

Keywords: Telecommunications, History, Interference, Power lines

Introduction

Optical fibre is ubiquitous in the Australian telecommunications network. One significant advantage is its immunity from electromagnetic interference due to the utilisation of light for the transmission of information, instead of electrical current.

Optical fibre is a relatively new technology, with commercial systems being pioneered in the 1980's. Telecommunication networks over the previous century had to rely on cables and wires to transmit information via electrical currents. Electromagnetic interference can generate noise in these conductors, particularly in proximity to high voltage alternating current power lines.

This interference would not have happened if direct current had been utilised for power reticulation, however by the turn of the twentieth century the world had standardised on alternating current. The contest between alternating and direct current for power distribution is well documented, suffice to say that alternating current was more cost effective, except in special applications such as the Basslink high voltage connection between Victoria and Tasmania.

It is often forgotten that telegraphic communication networks pre-dated electrical power networks in Australia by several decades. Both networks tended to follow the same easements along roads and railways and as the power line voltages increased, so did the electromagnetic interference.

The historic paper (Finlay, 1936) details a study undertaken in Tasmania between the Postmaster General's Department and the Hydro-Electric Commission into minimising

http://doi.org/10.18080/ajtde.v5n2.104

electromagnetic interference by coordinating the network roll-out planning between the two organisations.

The first half of the paper discusses the underlying principles in dealing with interference problems and gives examples of their application in practice. The second half focusses on the coordination of power and telephone systems and the practical field trial results.

The nature of power distribution networks is such that you can control interference from long runs of balanced three phase circuits using transposition; however in unbalanced spur lines, the interference can induce high residual voltages causing noise in telephone circuits that is difficult to control without coordination.

The paper concludes that great value can be achieved by coordinating the roll-out of power and telephone systems. In Tasmania, a Joint Committee of Engineers representing the Postmaster General's Department and the Hydro-Electric Commission was created to consider improvements to existing layouts and new construction proposals.

While researching this paper, I discovered a previous reader's hand written notes which neatly summarised the main conclusions. A copy of these notes has been included at the end of the historic paper for curiosity sake.

References

Finlay H. A. 1936. "Interference to Telephone Lines from High Voltage Transmission Lines", *Telecommunication Journal of Australia*, Vol. 1 No. 4 December 1936, pages 160-167.

The historical paper

http://doi.org/10.18080/ajtde.v5n2.104

INTERFERENCE TO TELEPHONE LINES FROM HIGH VOLTAGE TRANSMISSION LINES.

H. A. Finlay

The increasing development in the electrification of country areas throughout the Commonwealth has resulted in the erection of long extra-high and high voltage transmission lines operating from comparatively few generating sources. The creation of long parallels with adjacent communication circuits, with their severe inductive interference effect, has made the prevention of disturbance on telephone circuits an important and interesting study for Communication Engineers.

In Tasmania, progress in Rural Electrification has been considerable owing to the natural resources available. The Great Lake, located in the centre of the island at an altitude of 3,300 feet, is the chief source of electrical energy and from the Hydro-Electric generating stations located near the lake, 88 K.V. 3-phase 50-cycle Star-connected transmission lines radiate south, north-west, north and east to sub-stations, where isolated-neutral lines of 22 K.V. and 11 K.V. transformation Star-Delta YD (Delta on distribution side) distribute.

The power distribution systems are built along the highways and parallel, at small separation for the greater portion of their length, the main Interstate and Intrastate Trunk and Telegraph routes, as well as a considerable number of minor trunk and subscribers' routes.

Unfortunately, in the design, particularly of the 11 K.V. distribution system, many long single-phase (two-wire) branching circuits metallically connected to two wires of the three-phase distribution systems have been erected. This practice creates severe unbalances in the distribution system, causing high voltages on, and consequently considerable disturbance to, the neighbouring telephone circuits.

In what follows, it is proposed to indicate some of the more important underlying principles that have been established in dealing with the interference problem and to give some actual examples of their application in practice.

Inductive interference in communication systems due to power lines is directly proportional to the product of three simultaneously existing factors:—

"A"—The Inductive Coupling between the Power and Telegraph systems due to Mutual Capacitance and Inductance between the two systems.

"B"—The Disturbing Influence of High Voltage or Heavy Current Power systems.

"C"—The Susceptibility of Telephone lines.

To eliminate the interference it is necessary to reduce one of these factors to zero and conversely the increase of any factor results in a corresponding increase in the inductive effects. To secure complete freedom, the power transmission and telephone lines should be built on widely separated routes, but in Tasmania, where the conditions do not permit of a high degree of separation, consideration has had to be given to the reduction of all three factors. The principal components determining the magnitude of these factors are given hereunder in the order of their importance:—

"A"—The Inductive Coupling (Electrostatic and Electromagnetic) between Power and Telephone Systems depends on:

1. **Horizontal separation between the lines,** the inductance decreasing approximately as the cube of the separation or more accurately—

$$E \text{ (Induction)} \propto H^{-n}$$

where H is the horizontal separation and the exponent "n" depends on the power circuit configuration.

The lowest value of "n" is 2.3 for an isosceles triangle with altitude equal to 1.25 times the base; for equilateral triangle 2.6; vertical wires 2.9, and horizontal symmetrical 3.5. As doubling the separation will generally reduce induction to 1/8th, the importance of building routes with the greatest possible separation is evident.

2. **Length of Parallel,** the induction increasing in direct proportion to increase in the length of parallel.

3. **Spacing of Disturbing Power Conductors.** Increase of spacing tends to make the effect of the nearest power wire of a three-phase system more predominant, partly by decreasing its relative distance to the telephone wires and by increasing that of the other power wires and thereby reducing the neutralising effects of the latter. Induction generally increases proportionately to the spacing of the power wires.

4. **Spacing of the disturbed telephone wires.** Induction is directly proportional to the spacing between the two conductors of a metallic circuit. The 9 ins. spacing now adopted as standard for wires used for superposed carrier working results in a reduction of noise in lines exposed to inductive influence. In phantom circuits the noise produced in an untransposed section is three times that of the side circuit with the normal 14 ins. spacing.

5. **Height of Conductors.** As the voltage that is induced on a telephone wire from a neighbouring Power wire depends on the ratio of the Mutual Capacitance and the Capacitance to

http://doi.org/10.18080/ajtde.v5n2.104

ground of the telephone wire, increasing the height of both conductors tends to increase the induced potential and consequently the amount of interference.

"B"—The Disturbing Influence or Telephone Influence Factor of the Power System depends on:

1. **The Magnitude of the operating Voltages and Currents (balanced components),** the induction increasing directly with the voltage and currents.

2. **The Magnitude of the Residual Components of the operating Voltages and Currents.** The residual voltage or current of a power system is the vector sum of voltages or currents to ground of the conductors comprising the circuit, and is equivalent to a single-phase voltage impressed on, or current flowing in, a circuit consisting of the several power conductors in parallel and an earth return. Obviously, when the vector sums are zero the residuals are zero and a condition of Balanced Voltages and Currents exist.

The magnitude of the residual voltage depends on the unbalanced mutual and earth capacitances of the wires forming the power system and is particularly severe where long single-phase (two-wire) circuits are directly connected to the three-phase system, the phase wire which is not extended having, of course, a much lower capacity to ground. This practice results in the setting up of high residual voltages, both in the single-phase and three-phase portions of the power systems.

To overcome the severe residual effect due to this cause, it is necessary to—

(a) Electrically isolate the single-phase tap by means of an isolating transformer, thus eliminating the unbalanced capacitance to ground of the three-phase system and permitting the single-phase portion to become a balanced circuit with 180 deg. phase shift between wires, or

(b) To convert the single-phase tap to three-phase by erecting the third wire.

Where the power system is wholly three-phase, inequalities in the distances (capacitances) between power lines are of more importance than small inequalities in the distances (capacitances) to ground, and consequently various configurations or poling diagrams adopted for power lines have widely different characteristic residual voltage. For example, the horizontal configuration with unequal spacing of adjacent conductors gives rise to a much larger residual voltage than the equilateral triangular configuration.

Calculation of Residual Voltage.

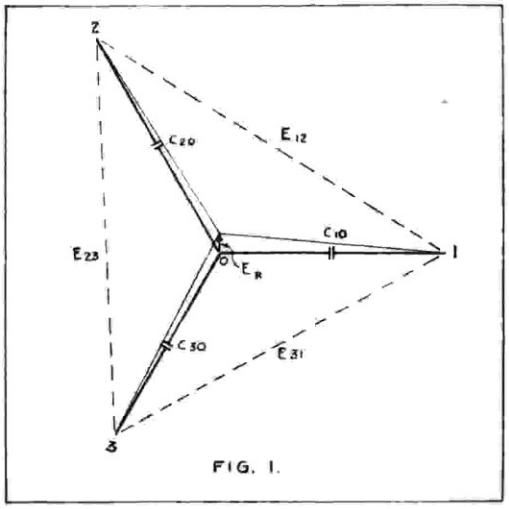


FIG. 1.

Figure 1 represents a three-phase system with unbalanced capacitances to ground and the formula for the calculation of the residual voltage is:—

$$E_R = -3E \frac{C_{10} + C_{20}\angle 120^\circ + C_{30}\angle 240^\circ}{C_{10} + C_{20} + C_{30}}$$

where E_R = Characteristic Residual Voltage.

$$E = \frac{E_{12}}{1.732}$$ (Voltage of each phase wire to ground)

C_{10}, C_{20}, etc., is the effective capacitance (Self and Mutual) of each phase wire to ground (O).

3. **Wave Form.** If a pure sine wave of fundamental frequency 50 cycles per second is propagated over power lines no noise would be heard in exposed telephone lines, owing to the high motional impedance of the telephone receiver and ear drum (together with the high capacitance reactance to this low frequency). Noise interference is due almost entirely to the odd multiples (harmonics) of the fundamental frequency into which the irregularities of the voltage and current waves may be analysed. The irregular wave forms may be caused by—

(a) Faulty design and construction of generators and motors.

(b) Distortion of voltage and current wave forms by faulty design and operation of transformers.

(c) Load unbalances of the power system.

The disturbing harmonics are divided into two groups—

(i) The odd triple series—3rd, 9th, 15th, etc.

http://doi.org/10.18080/ajtde.v5n2.104

(ii) The odd non-triple series—5th, 7th, 11th, etc.

In a balanced symmetrical three-phase system the voltage and current curves are separated by an angle of 120 deg. and their vector sum is zero, but if upon the fundamental the 3rd harmonic or any of its odd multiples is imposed, these harmonics will be coincident in phase and their effects will be additive and appear in the residual voltage.

On the other hand, the fundamental and odd non-triple harmonics can only appear in the residual voltage when they are unbalanced or unequal in magnitude. See Figure 2, which shows the magnitude of the residual 5th harmonic component of the voltage waves of the 11 K.V. power system radiating from the Bridgewater Sub-station.

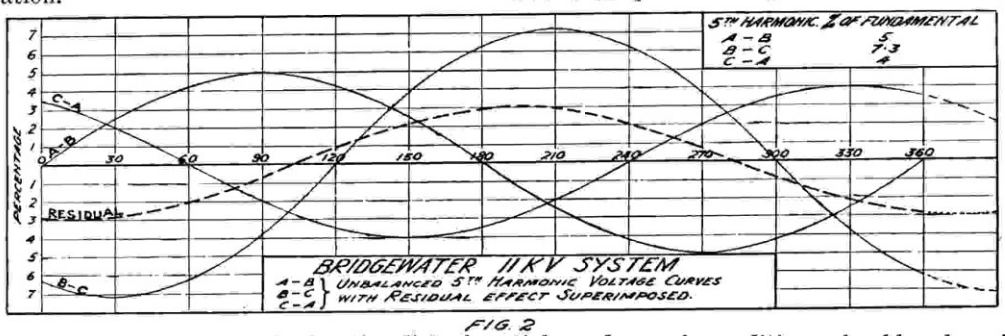


FIG. 2

As harmonics are responsible for the disturbance created in telephone circuits, special attention is given to their reduction by the improvement in generator and other power apparatus design by manufacturers, and modern specifications usually contain a clause limiting the maximum deviation from the sine curve to not more than 5 per cent.

The transformer connections can be used to suppress the 3rd harmonic and its multiples, a delta connection providing a closed path for the third harmonic current and its odd multiples, and so eliminating the corresponding harmonics from the flux and induced voltage wave.

For the suppression of a particularly troublesome harmonic, an anti-resonant circuit or wave trap can be designed for insertion in the generator earth.

"C"—The Susceptibility of the Telephone Line to Inductive Effects arises from the electrical unbalances in the two conductors forming the circuit or in the terminal apparatus connected to the line.

As earth circuit lines are wholly unbalanced with respect to earth, this type of telephone circuit cannot be retained when, due to development of Rural Electrification, portions of such circuits are brought within the sphere of influence of a High Voltage power system.

It is essential that the two sides of a telephone circuit be balanced both in series impedance and impedance to ground. The inductive susceptibility of a telephone circuit increases with—

1. **Insulation Resistance Unbalance** of each leg to earth. Circuits exposed to inductive influence and normally silent will become quite unworkable when a cracked insulator, wire contacting with a stay-wire or tree or contact by wet hands or oilskins of linemen causes localised insulation resistance unbalance.

On the other hand, humid and wet weather conditions bringing **uniformly distributed** leakage on the wires of each system does not increase the induction as is popularly supposed, and theoretically the reduction in capacitance and consequential longitudinally-induced potentials under such conditions should reduce inductive effects.

2. **Conductor Resistance Unbalance.** High resistance joints and different gauges or classes of wire will increase the intensity of disturbance. The permissible limit of conductor resistance unbalance is 2 ohms per 150 miles of circuit.

These two causes of unbalance can only be remedied by a high standard of telephone line construction, and a high grade of plant maintenance. Regular routine inspections to remove likely causes of unbalance are essential, and in this connection the principles for ensuring a high standard of maintenance of Trunk and Telegraph lines and the regular routine tests to determine such maintenance are laid down in General Engineering Circulars Nos. 8 and 17 respectively.

3. **Capacity Unbalance of telephone lines to ground.** In aerial transposed wires this component is comparatively unimportant, but in underground cables which are likely to be paralleled by extra high pressure lines, the balancing of the earth capacitances to the sheath of each wire is desirable to eliminate the interference due to electromagnetically induced currents flowing in the lead sheath.

4. **Terminal Apparatus.** It is essential that every unit of apparatus connected to a telephone

http://doi.org/10.18080/ajtde.v5n2.104

circuit be balanced within itself. When it is necessary to connect unbalanced apparatus or circuits (e.g., Earth circuit lines) to balanced circuits subjected to Inductive influence, a transformer to isolate the balanced from the unbalanced portion is always necessary.

CO-ORDINATION OF POWER AND TELEPHONE SYSTEMS—TRANSPOSITIONS.

Having dealt with the fundamental factors and their components in the causation of inductive interference, it is now proposed to discuss the fundamental principles underlying the inductive co-ordination of power and telephone systems.

Besides implying the adoption of proper maintenance and operating methods by both Power and Telephone administrations, so that the Inductive Influence of Power circuits on the one hand and the Inductive Susceptibility of the Telephone circuits on the other are reduced to the minimum possible, the term Inductive Co-ordination also refers to the best relative arrangement and location of transpositions in both circuits in order to diminish the inductive coupling and consequently the inductive interference.

Balanced Components of Voltage and Currents—the Effects of Transpositions.

In Figure 3, which represents a telephone line exposed to an untransposed three-phase power circuit for a distance of three miles, the effect of the nearest power wire is predominant and induces "longitudinal" voltages to ground along the wires of the telephone circuit. The voltage on each wire differs in magnitude, the larger, of course, being on the telephone wire nearest to the power lines, and this differential effect is known as "transverse" induction. The effect of inserting a transposition in the telephone circuit is to balance the induced potentials on each leg, neglecting phase change and attenuation, and if the series impedance and admittance to ground of each leg of the telephone circuit are equal, i.e., perfectly balanced, no circulating currents or noise will occur. Any imperfection in the balanced condition due, say, to unequal leakance in each leg or a high resistance joint will permit the longitudinally induced potential to cause current to flow through receivers and consequently generate noise.

In Figure 4, the effect of cutting in one barrel, i.e., a section wherein each power conductor occupies each of the conductor positions for the same distance—by means of two power transpositions (arranged right over left in this instance) and leaving the telephone circuit untransposed can be clearly seen. Here we see that over the section the induced voltage to

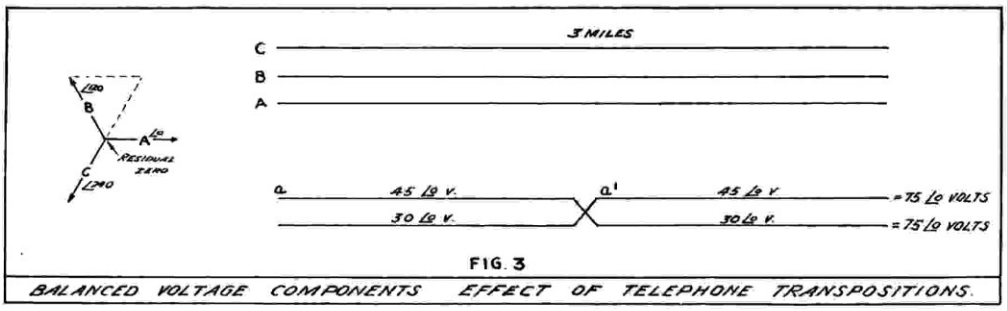


FIG. 3

BALANCED VOLTAGE COMPONENTS EFFECT OF TELEPHONE TRANSPOSITIONS.

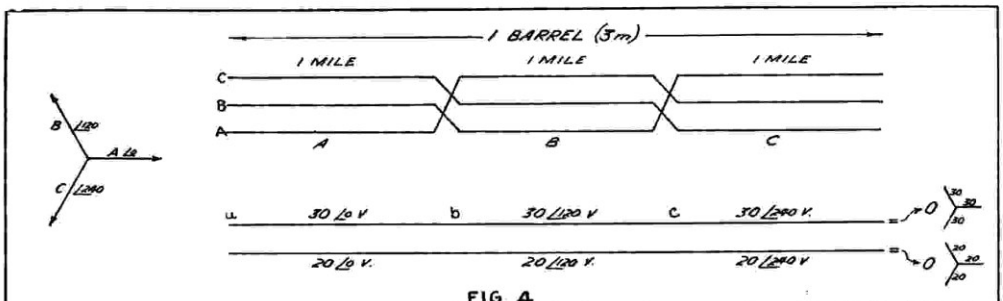


FIG. 4

BALANCED VOLTAGE COMPONENTS EFFECT OF POWER TRANSPOSITIONS.

http://doi.org/10.18080/ajtde.v5n2.104

ground is, by the neutralising effects of the three sections, reduced to zero on each leg, and also the difference of voltage between the two wires is zero.

In co-ordinated Power and Telephone transposition sections the telephone transpositions are located to balance the transverse induction from the nearest phase wire for each section of a barrel, and the power transpositions are used to reduce to zero in any one barrel the inductive effects from the balanced components of the power system.

Residual Components of Voltage and Current—the Effects of Transpositions.

The effect of residual voltage component of 11 K.V. three-phase system with two single-phase extensions is shown in Figure 5.

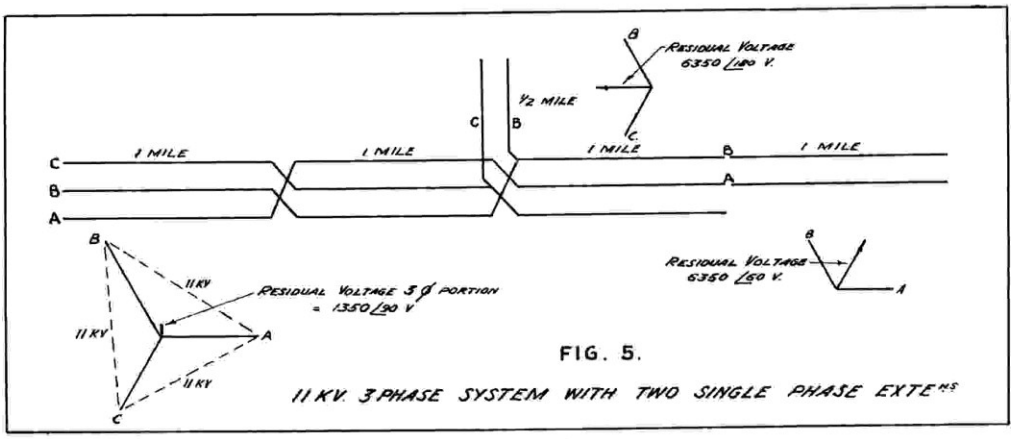


FIG. 5.

11 KV. 3 PHASE SYSTEM WITH TWO SINGLE PHASE EXTENS

The residual or unbalanced portion of the voltage persists throughout the three-phase and single-phase sections, and its effect can be represented by a single-wire ground-return power line with an E.M.F. equal to the residual voltage.

Figure 5 represents an 11 K.V. line (phase voltage to earth $\frac{11 \text{ K.V.}}{1.732} = 6{,}350$ volts) with two single-phase extensions, and is a typical example of an unbalanced power system. The vector diagram to the right indicates that the residual voltage on the single-phase extensions is equal to the phase to earth voltage of the system and located + 60 deg. from "A" and — 60 deg. from "B," i.e., midway between. As this voltage is wholly residual, i.e., totally unbalanced with respect to the neutral point, the insertion of a single-phase power transposition in such an unbalanced system will have no effect on parallelled circuits.

Figure 6 indicates the effect of the residual components of the power system illustrated in Figure 5 on a parallelled telephone circuit.

The residual voltage on the three-phase portion of the system is equal to 1350 volts and the approximate induced potentials on relative sections of the parallelled telephone wires are indicated.

The power transpositions in the three-phase portion neutralises the induced voltage to ground from the balanced components only, but cannot reduce the effect of the residuals. Telephone transpositions are effective against the induced potentials from residual voltages, but only equalise the voltages on each leg to ground.

Generally to obtain a high degree of transverse balance, it is of advantage to have numerous transpositions inserted in circuits erected on the pin positions located at each end of the telephone crossarms. Such well-transposed circuits would tend to shield adjacent circuits and also to reduce the secondary induction from themselves.

Owing to the effects of phase-shift and attenuation the necessity for more numerous transpositions increases with the increase of the frequency of the harmonics to be counteracted. By dividing the exposed circuit into smaller sections

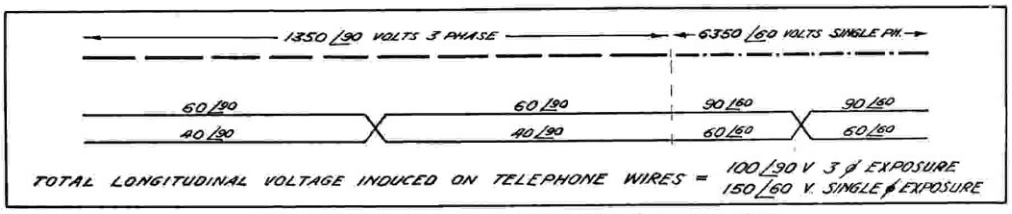


Fig. 6.—Effect of residual components of power system.

http://doi.org/10.18080/ajtde.v5n2.104

by transpositions, the induced voltages and currents in adjacent sections will have approximately the same phase and magnitude and accordingly will be more fully neutralised.

Unbalanced Exposures. A calculation of the unbalanced exposure or equivalent untransposed length by the formula hereunder for any given parallel gives a ready means of estimating the amount of induction to be expected and serves as a basis for computing the relative effects of any alterations to an existing power and telephone transpositions scheme:—

Type of Induction	Power Component	Length of Untransposed Exposure
Longitudinal	Residual	$a + b + c + a' + b' + c'$
	Balanced	$(a + a')\angle 0° + (b + b')\angle 120° + (c + c')\angle 240°$
Transverse	Residual	$(a + b + c) - (a' + b' + c')$
	Balanced	$(a - a')\angle 0° + (b - b')\angle 120° + (c - c')\angle 240°$

Note.—a and a', b and b', c and c' represent the lengths of each side respectively of a telephone circuit when subjected to exposure from, and A, B and C phase wires.

An actual example is given below. After the insertion of phantom transpositions in two 200C. lines between two centres approximately 90 miles apart, it was found that the phantom circuit could not be brought into commercial operation owing to heavy induction. The trouble was localised to a particular 7-mile section where the lines were parallelled by a 22 K.V. three-phase balanced power system at widely varying separations due to deviation of routes and curving roads. For the purpose of calculating the effective induction, sections having a greater separation than 11 yards were reduced to their equivalent length at this separation by formula $E = H^{-2.6}$.

22 K.V. Three-phase Parallel.

The calculations in the table below of equivalent unbalanced exposure indicated that the power transpositions were not suitably located to balance the longitudinal induction, and the telephone (phantom) transpositions were not suitably located with reference to power transpositions to provide balance against transverse induction.

An alteration calculated to reduce the longitudinal equivalent unbalanced exposure from 582 yards at half a chain separation, to the equivalent of 194 yards, by the interchanging of the B and C wires at Hydro poles Nos. 173 and 175 where wires were already terminated, was suggested to the Hydro-Electric Commission, but owing to some difficulty the alteration was performed at Pole 153 in lieu of 173. Calculations with the altered arrangement indicated that the unbalanced longitudinal induction was now the equivalent of 234 yards at 11 yards separation, approximately 40 per cent. of previous unbalance, and this compared closely with the observed reduction of noise, and this study made possible the handing over of the phantom circuit for telephone traffic.

Whereas the previous example was a case of induction primarily from the balanced components of a symmetrical three-phase system, the following case is of disturbance from a power system by no means so well balanced.

For some time past the Hobart-Granton portion of the main Trunk and Telegraph route has been periodically subjected to severe inductive disturbance, when, for pole renewals or other work, the Hydro-Electric Commission fed the parallelling 11 K.V. three-phase route south from the Bridgewater Sub-station instead of north from the Risdon Sub-station as normally operated.

An investigation into the cause of the trouble was recently undertaken and along sections parallelled by the Bridgewater system, where a

Section	Actual Length and Horizontal Separation	Computed equivalent lengths at ½ chain separation					
		a	a'	b	b'	c	c'
H.T. Poles:							
222 to 207	3,484 yds. at ½-chain separation	—	—	340	660	—	—
207 to 190		696	388	—	—	—	—
190 to 173		—	—	—	—	800	600
173 to 153	1,800 yds. at 2 chains 8 yds. separation	—	—	—	—	19	12
152 to 134	1,420 yds. at 1 chain 15 yds. separation	37	24	—	—	—	—
134 to 126	Not considered. 1,300 yds. at ¼ mile separation	0	0	—	—	—	—
125 to 120	400 yds. at 14½ yds. separation	—	—	—	—	195	—
120 to 107	Not considered. 1 mile at ¼ mile separation	—	—	—	0	0	—
106 to 90	1,400 yds. at 2 chains separation	15	23	—	—	—	—
87 to 76	1,200 yds. at 2 chains 15 yds. separation	—	—	—	—	9	9
		748	**435**	**340**	**660**	**1023**	**621**

Longitudinal (Balanced components): $1183 \angle 0°$; $1000 \angle 120°$; $1644 \angle 240°$.

Equivalent unbalanced exposure (Longitudinal) = 582 yds. @ (½ chain).

http://doi.org/10.18080/ajtde.v5n2.104

reasonable degree of co-ordination of the three-phase power and telephone transpositions existed, tests of the induced noise and voltage to ground on well-balanced telephone lines were made.

The fairly high induced longitudinal voltages to ground observed on telephone circuits indicated that the magnitude of the residual voltage component of the power system was high, and the residual effect of the predominant unbalanced harmonic, the 5th, of the voltage wave form was up to 3 per cent. of the fundamental. See Figure 2. A check-up of all power transpositions, connections of single-phase taps, lengths of cables and aerial wires was made over the whole extent of the high voltage system.

This 11 K.V. system furnishes light and power to an area of 1,000 square miles of country and consists of 65 route miles of three-phase and 80 route miles of single-phase branches. A special feature is that at the end of one 15-mile section of single-phase aerial wire extension, 800 yards of three-core, sector-shaped conductor, paper-insulated, oil-impregnated, lead-covered and armoured cable is laid to an Aerodrome, two cores only being used.

The effective capacitance of each core to sheath of the cable above-mentioned is given by the formula:—

$$C = \left(1.2 \frac{0.144}{\log_{10} \frac{d_2}{2r}}\right) 10^{-6} \text{ farads per mile}$$

where r = conductor radius.
d_2 = $2r$ + Insulation thickness between conductors + Insulation thickness between conductor and sheath.

The effective capacitance of one mile of an aerial wire forming one of a system of similar conductors is given by:

$$C = \left(0.03883 \frac{1}{\log_{10} \frac{2h}{r}}\right) 10^{-6} \text{ farads per mile.}$$

where h = height above ground.
r = radius of wire in similar units.

By calculation, the capacity of the High Voltage cable is 27 times that of an equal length of the aerial wire to which it is connected.

A copy of the computations and a suggested re-arrangement of the connection of certain single-phase taps in order to reduce the residual voltage was supplied to the Hydro-Electric Commission and with the co-operation of the Controlling Power Engineers the suggested re-arrangement was made.

The effect of the action was then specially tested under the conditions of feeding the Hobart-Granton parallel from the Bridgewater Sub-station and the results obtained were good, the total noise units being of the order of one-third of previous readings under similar conditions and little different from those recorded under normal conditions.

The following example shows the high permanent potential which can be induced on telephone lines parallelled by a single-phase extension of an 11 K.V. three-phase system.

From the underground cable terminal pole located 447 yards from an Exchange, an aerial wire subscribers' route is parallelled by a single-phase extension for a distance of 3.5 miles at an average lateral separation of 20 feet. The average height of the telephone wires is 16 feet and the power wires 28 feet.

The current to earth measured on three subscribers' lines traversing the full distance gave values of 3.2, 2.75 and 2.6 milliamps respectively, using an Elliott Milliammeter (Dynamometer principle range 0-15) of 1460 ohms resistance. A 1000 ohms magneto bell when connected between one leg of a subscriber's line and earth was caused to ring continuously by the induced current at fundamental frequency (50 cycles). The measured voltage induced between the lines and earth varied with the impedances of the instruments—one of 11,500 ohms resistance indicating 30 volts and another of higher resistance 75 volts. A diagrammatic representation of this test is shown in Figure 7.

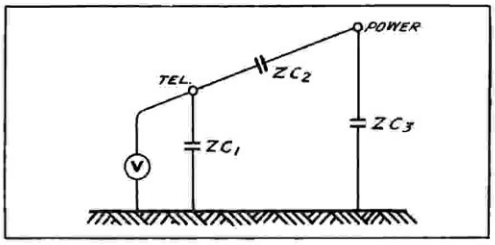


FIG 7

From the average dimensions of the parallel the earth and mutual capacitances and current flowing between power and telephone wires and earth were calculated and the induced voltage for the aerial wire portion disconnected from cable was computed to be 330 volts, and when connected to the Exchange via the cable 200 volts, also the current flowing from the telephone wires to earth was calculated to be 3.9 m.a's. The residual voltage of the single-phase 11 K.V. power line = 6350 volts and the calculations are accurate to within 10 per cent.

The electrostatically induced voltage is independent of the length of parallel and depends on the ratio of the impedances ZC_1 and XC_2, the connection of the comparatively low impedance voltmeter shunting ZC_1 immediately reduces the ratio and accordingly the voltage observed. The effect of an increase of parallellism, however, increases the mutual capacitance and accordingly the value of the current flowing between the two systems and with sufficient length of exposure to such unbalanced power lines, severe

http://doi.org/10.18080/ajtde.v5n2.104

damage to apparatus, or possible danger to personnel from electric shock, may occur.

To eliminate these effects at their source, it is necessary to secure the co-operation of the Electric Authority to erect the third wire or insert an isolating transformer between the single-phase tap and the three-phase system. Failing the above remedies, the induced voltage can be:

(a) Drained by connecting the primary of transformer with centre point earthed, across the subscriber's line, the secondary coil being left unconnected.

(b) Neutralised by connecting a neutralising transformer with associated auxiliary wire at the mid-point of the exposure.

Noise Measurements. In Tasmania, for the determination of noise levels on the various circuits and to observe the relative disturbing effects before and after the completion of rearrangements in power or telephone systems, a shunted receiver testing set is used. This set includes a resistance of 400 ohms in series with a variable resistance (ranging from 0.1 ohms to 111 ohms) which shunts a Bell receiver of 200 ohms impedance. The terminals of the set are connected to the noisy line and the variable resistance is decreased until noise in the receiver is just inaudible and the value of shunt noted. The shunt values are calibrated in Noise Units from 3,600 noise units corresponding to shunt of 1.0 ohms, to 50 units corresponding to 50 ohms.

As the values obtained are dependent on the sensitivity of the observer's hearing, differing results being obtained by different observers, it is necessary to have the observations taken with the same observer for a true comparative value to be obtained.

Also the values obtained are affected by any extraneous noises in the vicinity when tests are being made. Moreover, no allowance is made in the noise shunt of the relative interfering effects of the various frequency components of induced noise.

Experiments have shown that the relative interfering effects of the same power at differing frequencies is not the same, the interference being approximately proportional to the cube of the frequency up to 800 cycles per second, then rises slowly to a maximum at 1050 cycles.

Any noise measuring apparatus should, therefore, include some means of "weighting" the different frequency components so that the resultant measurement will be an indication of the detrimental effects of the noise on the intelligibility of speech over the circuit. This can be done by including a weighting network with an attenuation characteristic that is complementary to the relative interfering effects of the frequencies of the harmonics present.

Abnormal Conditions. The effect of a ground occurring on one-phase wire of an isolated neutral power system is illustrated in Figure 8.

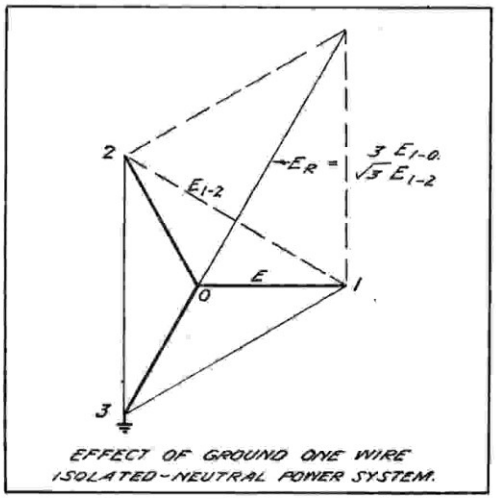


FIG. 8.

The residual voltage E_R which prior to the faulty condition may have been nil rises suddenly to three times the normal phase voltage to ground, and equals 1.732 times the phase to phase voltage, and this abnormally high residual voltage extends throughout the power system.

During switching operations, large residual voltages and currents may exist momentarily due to the several conductors not being energised or de-energised at the same instant. Also open circuits caused by a broken wire or short-circuits will give rise to heavy surges.

Accompanying the above abnormal conditions, severe acoustic shock may be experienced by telephone users or operating staff, and any such cases of shock must be closely followed up to ascertain possible causes and promote remedial measures. One measure at present under test is the use of neon gas-filled arresters, which have a low striking voltage, associated with a step-up transformer. The Hydro-Electric Commission co-operates to the extent of advising this Department, as far as possible, when switching operations are contemplated so that shocks from this cause can be reduced to a minimum.

The severity of such residual effects are minimised by the adoption of the grounded neutral star-connected system, which provides special measures to eliminate the effect of the third harmonic, and this method of high voltage power distribution is now generally favoured.

Conclusion. The value of co-operation is obvious and in Tasmania a Joint Committee of Engineers representing the Hydro-Electric Commission and the Postmaster-General's Department meets monthly to consider improvements to existing layouts and new construction proposals.

http://doi.org/10.18080/ajtde.v5n2.104

No 18

Inductive interference due to power lines is dependent on

1. Inductive coupling between the two systems due to mutual capacitance + induction.
2. Disturbing influence of high voltages + heavy currents.
3. Susceptibility of the telephones lines.

Inductive Coupling depends on

a) Horiz. distance between the two systems
(doubling separation generally reduces induction by about ½)

b) Length of parallel, the induction increasing in direct proportion to increase in the length of the parallels.

c) Spacing of the wires; reducing the spacing in either system reduces the induction.

d) Height of wires; as the induced voltage depends on ratio of mutual capacitance + capacitance to ground of the telephone wire, increasing height of either system increases interference.

http://doi.org/10.18080/ajtde.v5n2.104

An Evaluation and Enhancement of a Novel IoT Joining Protocol

Tyler Nicholas Edward Steane
RMIT University

P J Radcliffe
RMIT University

Abstract: The ability to securely join IoT Devices to Wi-Fi networks is an ongoing area of research. This paper describes how Nasrin & Radcliffe's theoretical "novel minimalist IoT network joining protocol" has been mapped to real world hardware and implemented using the Android operating system. For the first time the theory is proven to be practically viable but it is also shown that the user interface is not sufficiently simple for the everyday user. This paper proposes and implements a new user interface paradigm that dramatically simplifies the process and makes the joining process accessible to a much larger range of users. For intensely cost-sensitive applications an alternative process is proposed that has the possibility of even further simplifying the user experience. Finally, the compatibility of the protocol with a variety of operating systems is assessed.

Keywords: Internet of Things, IoT, home automation, smart home, distributed discovery protocol

Introduction

The Internet of Things (IoT) is a field of rapidly growing interest with promising applications in home automation (Gubbi et al , Buyya, Marusic, & Palaniswami, 2013). Solutions are being offered at both a research level and a commercial level. These solutions are encumbered by high expense, complicated interfaces and poor security (Greichen, 1992) and these limitations all have an effect on the user's experience (Chang et al, Dong, & Sun, 2014). Creating a simple and positive user experience is essential to the success of IoT devices, particularly in the area of home automation where everyday consumers want to install and use an IoT device.

One example of a commercially available IoT device in the Home automation space is the 'Nest', a smart thermostat for the home with a rotating ring and LCD display for an interface (Nest, 2015). While this is sufficient for simple tasks it is cumbersome for entering the Wi-Fi password. Many devices only need such an interface to connect to a network, thus an

http://doi.org/10.18080/ajtde.v5n2.92

interface is often more complex than the other functions of the device and thereby such an interface dramatically inflates the price of the device.

To address these issues a novel network joining protocol specifically for IoT home automation has been proposed in a theoretical fashion by Nasrin and Radcliffe (2016). This work made basic suggestions for the protocol's implementation as well as an interface to accompany its use. This approach would reduce cost and interface complexity by using pre-configured Wi-Fi credentials and the ubiquitous smartphone with its wireless routing capabilities or hotspot.

While this was a promising protocol it had not been physically implemented on existing hardware. This paper details how the protocol has been mapped to the capabilities of the Android OS and then implemented using real Android hardware. The interface will be shown to be insufficient for every day users and so it will also be shown how significant enhancements to the interface have been implemented to improve the user experience.

This paper is organised as follows: Section II reviews existing work in this area. Section III briefly outlines the approach taken by Nasrin and Radcliffe. Section IV details how Nasrin and Radcliffe's protocol has been mapped and implemented on real hardware. Section V will assess and enhance Nasrin's interface. Section VI will suggest alternate enhancements that might be explored. Section VII provides an assessment of Nasrin and Radcliffe's protocol beyond the Android OS. Finally, Section VIII details the areas of focus for future work on this protocol and its implementation.

Existing Work

Four main approaches have been seen in the design of home automation devices, they are: Dedicated IO; Bridge; Central Controller; and a Minimalist approach.

Dedicated I/O

In this approach, additional IO hardware is used solely for the purpose of securely joining the local network. Dedicated IO protocols are often different from the primary communications protocol (usually Wi-Fi); for example, NFC, Bluetooth, or a fully featured user interface (with a screen and keyboard or similar input control) have been included in IoT devices purely for the one-off joining event.

Chen, Pan and Li (2012), for example implemented NFC Tags in devices, but still rely on a traditional network to establish a connection. The need for the user to physically move to each device makes the system, if anything, less usable than a mechanical switch. Others like Piyare and Tazil (2011) have used Bluetooth as a regular communications protocol but only

http://doi.org/10.18080/ajtde.v5n2.92

between a phone and central controller. The advantage of this was that end devices didn't need Bluetooth hardware and so are cheaper but comes with the added cost, complexity and inflexibility of a Central controller.

Dedicated IO works well for larger, more complicated appliances where additional IO hardware is trivial or already present. It provides some opportunity for the joining event to be streamlined and the user's experience may be marginally improved. However this approach does not scale well to simple devices where it would greatly increase size, complexity and cost. Compromises made using this approach can result in lower cost but questionable user experiences. The Nest example referred to earlier has a 1.75" round screen with a rotating ring. Rotation of the ring is used to scroll through options and depressing the ring will make a selection. This is an extremely compact approach sufficient for most regular operations but is inconvenient and error prone when used to input a Wi-Fi password to join the device to a network (Nest, 2015). This compromise would reduce the cost of the device compared to a larger keyboard and display but loses the advantage of a positive user experience. The overall cost of the device is much higher than a device with no IO hardware, and the user is still left with a clunky interface.

Bridge

The second approach to solving the network joining problem again uses additional hardware to translate between Wi-Fi and some other protocol, thus creating a bridge. Zigbee has been a popular protocol in this approach (Dou et al , Mei, Yanjuan, & Yan, 2009; Yan & Dan, 2010) as it has secure joining inherent in the protocol (Gomez & Paradells, 2010). Control of a device is thus maintained by a bridging device running a protocol like Zigbee connected to a local router. Other protocols like Insteon and Z-wave have been used; however, like Zigbee they ultimately rely on other protocols and infrastructures to be in place. These secondary protocols are simply bridging the IoT device over to another, primary, protocol to avoid building capabilities into the device to handle the primary protocol. Such an approach only serves to increase cost and complexity of IoT devices and the networks they inhabit (Gomez & Paradells, 2010).

Central Controller

A third approach to network joining uses a central controller to which all devices are physically wired. This physical connection offers an inherent security advantage and reduces the complexity of devices, but it sacrifices the flexibility offered by a wireless connection.

KNX is one of the more mature and successful protocols used in this approach. It is built from several well-established protocols and is tailored to situations such as home

http://doi.org/10.18080/ajtde.v5n2.92

automation (Lee & Hong, 2009). Most implementations employ a wired solution, but all approaches rely on a PC or server, as a central controller, being connected and powered on constantly. It also requires a considerable amount of processing power within each device (Zamora-Izquierdo et al., Santa & Gomez-Skarmeta, 2010). This means that as a solution it is very expensive: devices are expensive, central controllers are expensive, running costs are high, and adjustments are difficult and costly.

Such systems are not easily altered and users are quickly locked into a single product line, increasing costs. While the individual devices are simpler, the overall system's complexity is increased and its flexibility reduced, compromising any gains to the user experience.

Minimalist

Finally, Nasrin and Radcliffe's (2016) new minimalist approach has been proposed which uses Wi-Fi and the associated security protocols (e.g. WPA, WPA2) but, unlike the other approaches, removes the need for complex and cumbersome user interfaces. This approach uses the hotspot capabilities of a smartphone to establish an initial temporary, but secure connection to pass credentials for the local home network. This approach requires little or no adjustment to existing controllable devices and is extremely scalable to simpler devices such as mains switches.

The Minimalist approach, with its protocol promising uncompromised security at reduced cost and complexity, is the best approach. However, this approach has yet to be physically implemented; Nasrin has only offered a simple proof of concept which does not automate the hotspot function of the phone and requires attention to improve the user interface. The proposed interface requires a high technical competence from the user and does not streamline the process or maintain a positive user experience. With such modifications, this approach would be of immediate use to the IoT industry.

Nasrin and Radcliffe's Approach

Nasrin and Radcliffe's protocol is designed to connect IoT device to wireless networks while maintaining a high level of security. It is intended to reduce costs by only using the hardware already necessary for a wireless IoT device (Nasrin & Radcliffe, 2016).

Fig. 1 summarises the IoT joining protocol. Each wireless IoT device would be preconfigured with a unique Service Set Identifier (SSID) and password. The device can only be contacted on a network meeting these unique and secret settings. This can most easily be achieved by a smartphone acting in hotspot mode.

http://doi.org/10.18080/ajtde.v5n2.92

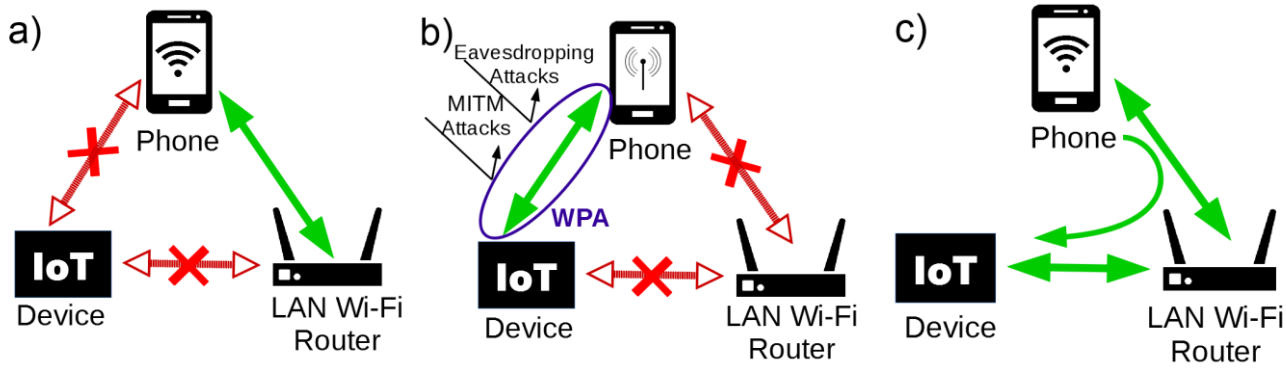


Figure 1: Nasrin and Radcliffe's Steps. a) Initial set-up, b) Hotspot mode transfer LAN details, c) Phone and IoT transfer to LAN

Typically, this temporary link would be secured by WPA2, allowing the SSID and password for the local Wi-Fi network to be passed to the IoT device in a secure manner. Once the IoT device receives the local Wi-Fi credentials it disconnects from the smartphone's hotspot and joins the local network permanently, along with the smartphone. All this is achieved without the need of extra hardware or complexity in the IoT device.

Thus far all that Nasrin has demonstrated is that when a smartphone and an IoT device are connected via the smartphone's hotspot the two can transmit data to one another using the User Datagram Protocol (UDP). Nasrin further envisages a more complete implementation of the proposed joining protocol, by the addition of two features: the first to automate the task of putting the smartphone in and out of Access Point (AP) mode and the second to automate the reconfiguration of the IoT and smartphone Wi-Fi credentials to join to the local Wi-Fi network. These functions need to be implemented to ensure the protocol is practically viable and this is one of the main purposes of this paper.

Mapping & Implementation

In assessing the capacity of current hardware to support Nasrin's protocol, four key functions needed to be achievable. Firstly, it needed to be possible to force a smartphone in and out of AP mode. Secondly, the Wi-Fi configuration of a smartphone needed to be programmatically configurable, in both Wi-Fi and AP mode. Thirdly, communications between an IoT device and a smartphone needed to be achievable. Finally, the Wi-Fi credentials of an IoT device needed to be programmatically configurable.

These key functions were mapped successfully to the Android operating system, while the IoT device was represented by a Raspberry Pi running Linux.

The following assessment was made to verify the compatibility of Nasrin's protocol with current hardware and to identify the tools necessary to implement the protocol:

http://doi.org/10.18080/ajtde.v5n2.92

Smartphone Wi-Fi Programmatic Initiation is possible using setWifiEnabled, a Boolean member of the android.net.wifi.WifiManager API class, which can toggle Wi-Fi on or off (Android Developers, 'WifiManager', 2016c). While the status of the Wi-Fi service can be determined using isWifiEnabled() or for more detail getWiFistate(). The getSystemService method (Android Developers, 'Context', 2016b) will return the WifiManager class to enable the above functions (Mendoza, 2012).

Programmatic Wi-Fi Configuration can be achieved using the addNetwork() member function of the WifiManager Class and configuring the WifiConfiguration class. By editing the WifiConfiguration class and populating the SSID and PreshareKey strings the Wi-Fi can be configured to the desired settings. In order to achieve configuration under AP mode Android 4.0 (API 14) or higher must be used earlier versions do not support configuration (Android Developers, '<uses-sdk>' ,2016a).

Communications were previously developed by Nasrin and Radcliffe (2016) using a UDP link to deliver packets between devices. This communication link is well understood and other possible approaches abound, however the existing work is sufficient.

IoT Wi-Fi Programmatic Configuration under Linux can be achieved by editing the configuration file, wpa_supplicant.conf (Malinen,2013) and then restating the service

All of these functions were successfully implemented and integrated into a functional Android application using the interface proposed by Nasrin and Radcliffe, shown in Fig. 2a-2b). The Raspberry Pi was used to represent an IoT device, in this case a simple Light which can be turned on and off, Fig. 2c). This was done by toggling a TRIAC using the GPIO pins on the Raspberry Pi, and connecting mains and a lamp to the TRIAC.

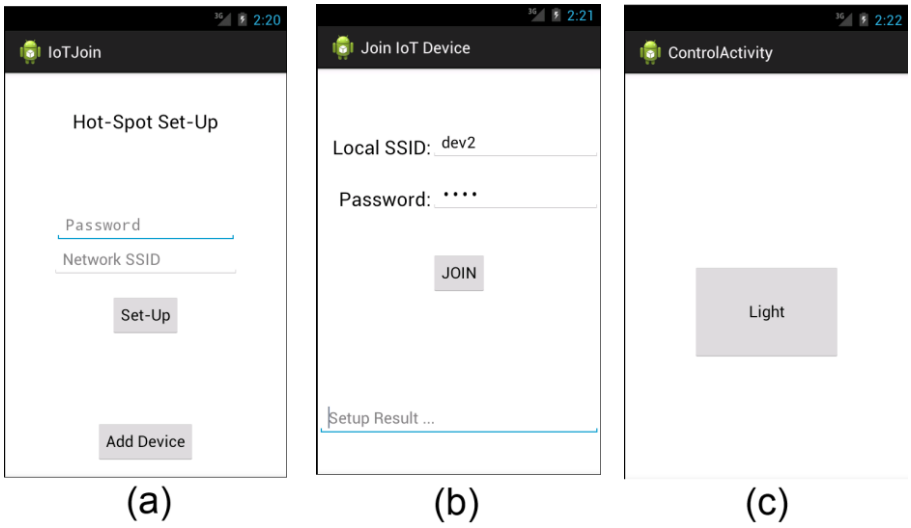


Figure 2 Implementation of Nasrin's interface in Android 4.0

http://doi.org/10.18080/ajtde.v5n2.92

This assessment for the mapping of Nasrin and Radcliffe's protocol has demonstrated the capacity of the Android OS to support the protocol. However further assessment is needed to ensure that other OS are capable of supporting the protocol.

Enhancing Nasrin's Interface

At this point limitations of Nasrin's suggested interface became apparent. It was an effective interface but a rather demanding one, requiring the user to be very familiar with their local Wi-Fi network's configuration. The interface also required typing of random SSID and passwords which increases the likelihood of user error and so creates a poor user experience. The interface further lacked any form of progress indication to reassure the user that the application was working on their requests and not simply waiting for further prompts or had frozen. Key enhancements implemented are discussed below.

Firstly, an automated method for entering the preconfigured credentials of the IoT device was developed using QR codes. Instead of the manufacturer providing a unique written SSID and password they would supply a unique QR code which the user would scan thus saving on typing, errors and time. The successful implementation is shown in Fig. 3.

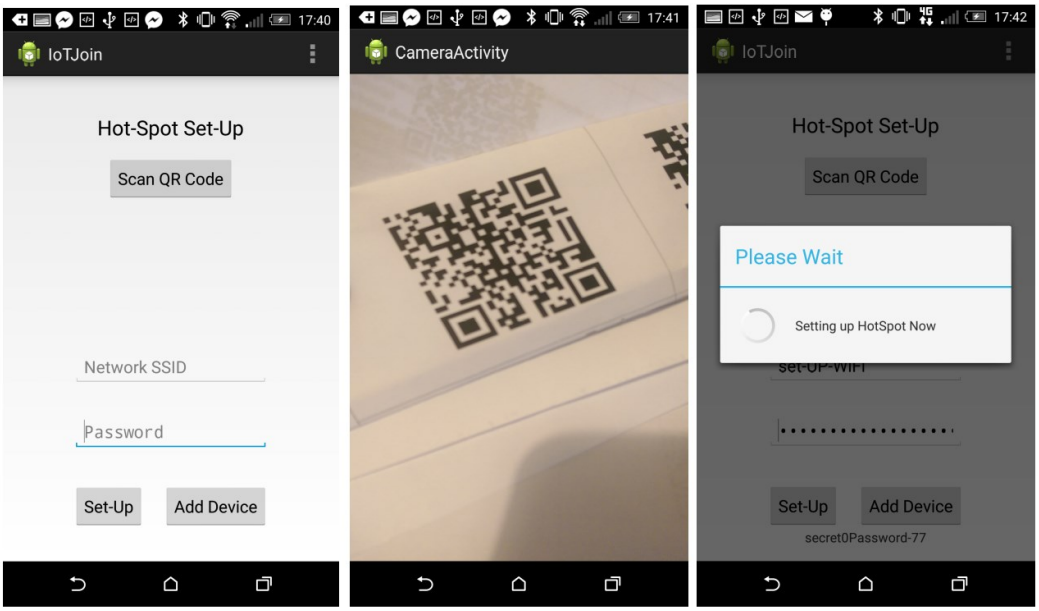


Figure 3 Hotspot setup using QR code scanning

http://doi.org/10.18080/ajtde.v5n2.92

Secondly, users are more accustomed to joining Wi-Fi networks by selecting the desired SSID from a list. This approach has been integrated into the joining application, shown in Fig. 4. Now users only need to be able to identify their SSID (not type it) and enter the password. This provides a familiar process for the user and requires only minimal familiarity with the home Wi-Fi.

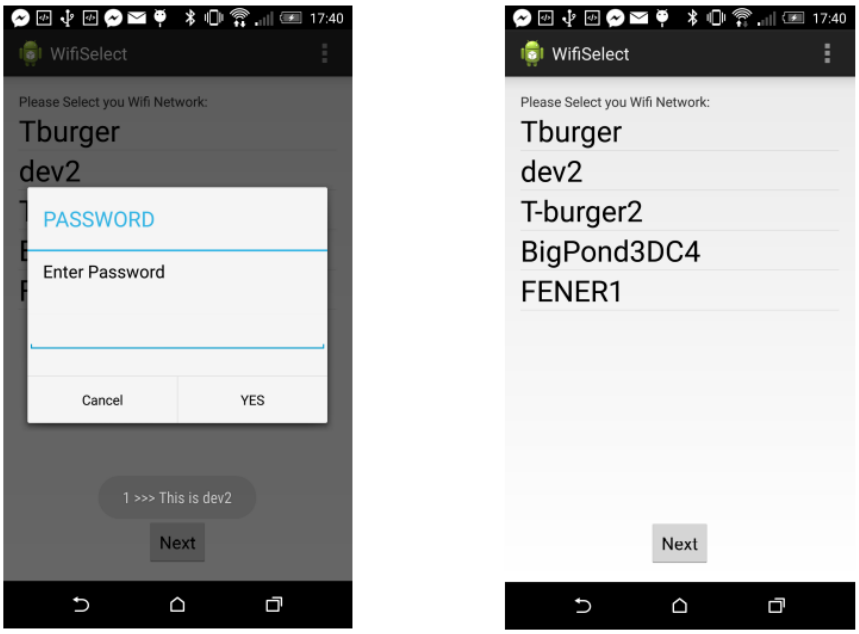


Figure 4 Available Wi-Fi list and joining dialogue

Thirdly, Loading screens are a simple but effective way of reassuring users that the application is working and not simply sitting idle. They can be used to inform of wait times and if something is not configured correctly. If the user's expectations are set early, longer wait times have less impact on the user's experience, whereas unexpected inactivity quickly leave users uncertain and concerned, quickly eroding the user experience.

Fig. 5 shows the complete signalling detail between all actors for the final implementation. The drawing of Fig. 5 caused us to uncover an omitted but important use-case: how can the IoT device be removed from a network and joined to another? One solution is to include a Wi-Fi based command to leave the network but this presupposes that the original network is available to allow this transaction. A better solution can be borrowed from nearly all routers: a very inexpensive pinhole factory reset switch that sets the IoT device back to its unique SSID and password.

http://doi.org/10.18080/ajtde.v5n2.92

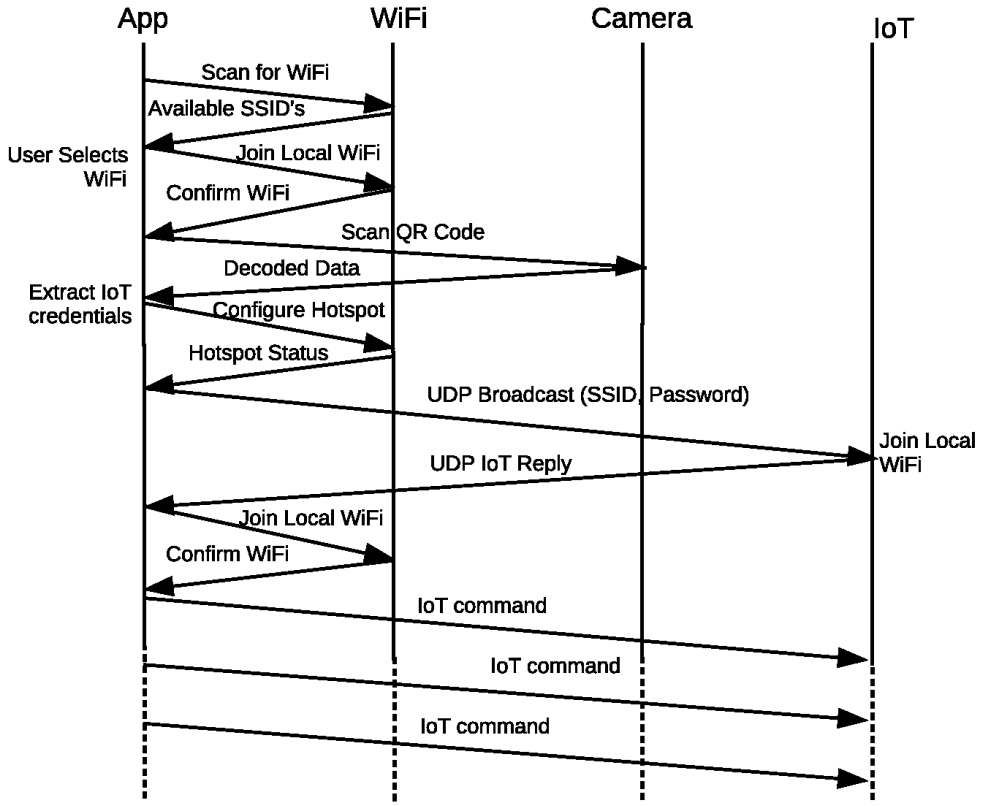


Figure 5 A bounce diagram for a successful connection

Simpler Again & More Secure

Proving that Nasrin & Radcliffe's network joining protocol worked in practice, and making significant enhancements, has helped make the protocol more user-friendly – but is it possible to make the user experience even simpler? While cost has ruled out extra hardware solutions such as adding NFC, these implementations can result in a very simple interface: the user simply selects their own LAN, enters the password, and hits a "join" button.

In order to preserve the minimalist approach of Nasrin and Radcliffe's protocol is it possible to make use of the hardware already inherent in IoT devices to simplify the user's experience? IoT devices often have extra hardware of very low cost, such as an LED indicator or a buzzer, which are under the control of a microprocessor.

It may be possible to use these hardware components to simplify the joining process as much as NFC has achieved. Using Nasrin & Radcliffe's protocol these LEDs or buzzers could securely send the IoT's unique SSID and password to the smartphone thus eliminating the QR code stage. This could reduce costs and simplify the users' experience.

Instead of scanning a QR code users would start up the IoT device which would begin flashing (for the LED) or playing (for the buzzer) a signal encoded with the local SSID and password. Users could then capture this signal from a smartphone application, whether by

http://doi.org/10.18080/ajtde.v5n2.92

pointing the camera at the LED or by holding the microphone to the buzzer. This removes the need for QR codes, eliminating the manufacturers cost of generation, printing and packaging. Furthermore, users won't need to keep track of QR codes and which devices they belong to.

This approach adds an extra level of security as the joining device must be present – not just within range of Wi-Fi, but also within range of the buzzer or LED. Thus, the IoT device will be able to verify that it is communicating with a user within a close enough proximity to likely be the owner (or at least possessor) of it.

In Nasrin and Radcliffe's original design of the protocol, if a malicious user happened to know the initial IoT SSID and password and was within Wi-Fi range of the IoT at the time of the devices being powered on prior to joining, they might be able to join the device first or capture the home SSID and password. However, with the use of auxiliary hardware the SSID and password can only be retrieved by a much closer proximity. To prevent malicious users guessing the Wi-Fi credential or predicting them based on manufacturers' patterns, the IoT device could further verify the device either by randomly generating the password or after connecting to the hotspot by requesting a second random pass code which would be encoded on the LED or buzzer. The process by which these passwords are generated need not be overly complex so as to demand significant overheads, and need only avoid basic predictability by avoiding patterns related to the manufacture. The main purpose and value is that it is not a static password.

These options, if feasible, have the potential not only to simplify the users experience but also to reduce the cost of IoT devices as well as to enhance the security of the joining event by verifying the user's proximity to the IoT device.

LED Solution

An LED implementation would require the user to hold the smartphone's camera over a flashing LED to obtain the unique SSID and password for the IoT device.

An IoT device can modulate an LED at high speed but a smartphone can only monitor such an LED using video capture at frame rates as low as 12 Hz up to 60 Hz or more depending on the smartphone. Using the Nyquist sampling criteria this means the data rate at best will be between 6 Hz and 30Hz. While an SSID can be quite short, perhaps only 2 letters, a password should be much longer. For WPA2 it has been suggested that a password be a minimum of 12 random characters which equates to 78 bits of entropy (Farik & Ali, 2015a; 2015b). Thus, a minimum of 14 characters, 2 for SSID and 12 for a password, each of 7 bits is required; 98 bits. At a 6 Hz sample rate a user would thus have to hold the camera over the

http://doi.org/10.18080/ajtde.v5n2.92

IoT LED for approximately 17 seconds, or 4 seconds for a high-speed smartphone. These times are based on calculations for a baseband signal, they may be reduced by more complex modulation and encoding but this would come at the cost of increased computational overheads which may not be viable for constrained simple devices. It would be quite a challenge to design a user interface that would help a user cope with these time lengths. Scanning a QR code would appear to be a much simpler solution for the user.

Buzzer Solution

The majority of buzzers are of a self-resonant type which buzz at a predetermined frequency when supplied with a DC voltage. In terms of modulation this implies that amplitude modulation (AM) is the only viable modulation technique. When a small buzzer is turned on and off the attack and decay times are typically 2-3 milliseconds which suggest a data rate of at least 100 Hz is possible. A smartphone can capture and record audio information at speeds from 3 kHz up to 44 kHz depending on the quality of the smartphone. Assuming a low data rate of only 100 bits per second this means that the smartphone could capture the IoT SSID and password in around one second given that AM modulation is the only viable modulation method.

While this solution seems viable the annoyance value of a buzzer must be taken into consideration. A continually operating buzzer is likely to annoy a user and erode the user experience significantly. Perhaps the factory reset switch previously mentioned could be used to play out the SSID and password several times and then stop.

Compatibility

Having now implemented Nasrin and Radcliffe's protocol we are in a position to accurately understand the requirements of the protocol and to consider the compatibility of the protocol with other operating systems.

Under Android, API's are available to grant access to control the Wi-Fi Hardware and as such the application manifest must declare that this application will be granted access to these settings. Can other mobile operating systems grant such access or is such access blocked? Most notably, Apple's iOS does not allow developers to programmatically configure the Hotspot mode, nor even to switch it on or off; some access is available to Wi-Fi connectivity and configuration but it is very limited and tightly controlled (Apple Developer Documentation, "CoreWLAN", 2017). While this has been confirmed by a careful assessment of the developer's documentation, it is also widely confirmed by the official and unofficial developer forums.

http://doi.org/10.18080/ajtde.v5n2.92

While Android has a large share of the market at 86.8% globally it is by no means without competition (IDC, 2017). However, in countries like Australia Android leads but only with 52.3% of the market followed closely by Apple with 44.9% (Kantar Worldpanel ComTech, 2017). Similarly, in Great Britain Android has 50.6% of the market share compared to 47.6% held by Apple and in the USA Android holds 54.4% with Apple at 44.4% (Kantar Worldpanel ComTech, 2017). Given these market figures, any protocol should work with both Android and iOS.

Without the ability to programmatically interface with the Wi-Fi hardware under iOS, a smooth user experience is not possible. However, it would not even be possible to implement Nasrin and Radcliffe's protocol with manual Wi-Fi configuration. This is because users cannot configure the Hotspot SSID, which is created based on the devices name.

Furthermore, not all Android devices are equipped with hotspot capabilities or they have been disabled. It is therefore, on the whole, not reasonable to assume that anyone wanting to configure a home automation system will have access to compatible hardware. Thus, it would seem that while Nasrin and Radcliffe's Protocol can be practically realised and with some enhancements it can be made user-friendly, it ultimately fails to address the availability of hardware to the average user.

The best solution would be a universal joining protocol compatible with any devices that supports Wi-Fi. Currently Nasrin and Radcliffe's protocol is not able to support this and so further work is needed to make it a viable solution for all users in the DIY home automation market.

Future Work

The enhanced version of Nasrin and Radcliffe's protocol is very viable but the use of an LED or buzzer to send the IoT SSID and password to the smartphone does have the potential to further simplify the user's experience. Developing an LED-based solution is a real challenge requiring an outstanding user interface or some very novel sampling methods. The buzzer solution appears more viable and is worth further investigation.

Nasrin and Radcliffe's protocol has been successfully implemented for the first time, and this has allowed an assessment of its compatibility with common devices. This assessment raises concerns around the limited compatibility of the protocol as it is essentially only compatible with Android devices with hotspot capabilities. Further work will investigate alterations to the protocol to widen its range of compatibility to include at least iOS. A joining protocol for home automation devices must be as widely compatible as is reasonably possible.

http://doi.org/10.18080/ajtde.v5n2.92

Conclusion

It has been demonstrated by a consideration of the literature that, in theory, Nasrin and Radcliffe's protocol is both a novel and superior network joining protocol for IoT devices. Careful analysis has shown that the protocol can be mapped to the capabilities of the Android operating system. The physical implementation of Nasrin and Radcliffe's protocol, using an Android smartphone and a Raspberry Pi, has proven that the protocol is practically viable. It was discovered that while Nasrin's user interface was functional it was not friendly for the everyday user. This paper proposed and implemented three new user interface features that made the network joining process notably easier for the average user.

Many simple IoT devices have an LED or buzzer under control of a microprocessor. This paper has analysed the potential for these devices to replace the QR code portion of the joining protocol, and concluded that such an approach is viable and worthy of future research.

Finally, the compatibility of Nasrin and Radcliffe's protocol has been determined to be limited to Android devices with hotspot capabilities. While this limitation still includes many devices, it is still incompatible with many popular devices notably those running iOS. This provides an additional area for further research.

Acknowledgements

Thanks to Kai Xu for his work testing LED's and Buzzers.

This research was supported by an Australian Government Research Training Program Scholarship.

References

Android Developers. (2016a). <uses-sdk>. Retrieved June 13, 2016, from https://developer.android.com/guide/topics/manifest/uses-sdk-element.html#uses

Android Developers. (2016b). Context. Retrieved June 13, 2016, from https://developer.android.com/reference/android/content/Context.html

Android Developers. (2016c). WifiManager. Retrieved June 13, 2016, from https://developer.android.com/reference/android/net/wifi/WifiManager.html

Apple Developer Documentation. (2017). CoreWLAN. Retrieved March 17, 2017, from https://developer.apple.com/reference/corewlan

Chang, Y; Dong, X; Sun, W. (2014). Influence of characteristics of the Internet of Things on consumer purchase intention. *Social Behavior and Personality*, *42*(2), 321–330. Available at https://doi.org/10.2224/sbp.2014.42.2.321

Chen, L; Pan, G; Li, S. (2012). Touch-driven interaction via an NFC-enabled smartphone. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (pp. 504–506). Available at https://doi.org/10.1109/PerComW.2012.6197548

Dou, N; Mei, Y; Yanjuan, Z; Yan, Z. (2009). The Networking Technology within Smart Home System - ZigBee Technology. In *International Forum on Computer Science-Technology and Applications, 2009. IFCSTA '09* (Vol. 2, pp. 29–33). Available at https://doi.org/10.1109/IFCSTA.2009.129

Farik, M; Ali, A. S. (2015a). Algorithm To Ensure And Enforce Brute-Force Attack-Resilient Password In Routers. *International Journal of Technology Enhancements and Emerging Engineering Research*, *4*(10), 184–188.

Farik, M; Ali, A. S. (2015b). Analysis Of Default Passwords In Routers Against Brute-Force Attack. *International Journal of Technology Enhancements and Emerging Engineering Research*, *4*(9), 341–345.

Gomez, C; Paradells, J. (2010). Wireless home automation networks: A survey of architectures and technologies. *IEEE Communications Magazine*, *48*(6), 92–101. Available at https://doi.org/10.1109/MCOM.2010.5473869

Greichen, J. J. (1992). Value based home automation for todays' market. *IEEE Transactions on Consumer Electronics*, *38*(3), XXXIV–XXXVIII. Available at https://doi.org/10.1109/30.156666

Gubbi, J; Buyya, R; Marusic, S; Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, *29*(7), 1645–1660. Available at https://doi.org/10.1016/j.future.2013.01.010

IDC. (2017). IDC: Smartphone OS Market Share. (2017). Retrieved February 24, 2017, from http://www.idc.com/promo/smartphone-market-share/os

Kantar Worldpanel ComTech. (2017). Smartphone OS sales market share. Retrieved February 24, 2017, from https://www.kantarworldpanel.com/smartphone-os-market-share/

Lee, W. S; Hong, S. H. (2009). Implementation of a KNX-ZigBee gateway for home automation. In *2009 IEEE 13th International Symposium on Consumer Electronics* (pp. 545–549). Available at https://doi.org/10.1109/ISCE.2009.5156866

Malinen, J. (2013) Linux WPA Supplicant (IEEE 802.1X, WPA, WPA2, RSN, IEEE 802.11i). Retrieved June 13, 2016, from http://w1.fi/wpa_supplicant/

Mendoza, A. J. (2012). Tutorial For Android: Turn off, Turn on wifi in android using code tutorial. Retrieved June 13, 2016 , from http://www.tutorialforandroid.com/2009/10/turn-off-turn-on-wifi-in-android-using.html

Nasrin, S; Radcliffe, P. J. (2016*). A Novel Three Stage Network Joining Protocol for Internet of Things based Home Automation Systems. Computer Communication & Collaboration, 4(3)*, (pp. 1-11).

Nest. (2015). Nest Protect and Nest Cam support. (2015). Retrieved June 13, 2016, from https://nest.com/support/article/A-step-by-step-guide-to-setup-on-the-Nest-Learning-Thermostat

Piyare, R; Tazil, M. (2011). Bluetooth based home automation system using cell phone. In *2011 IEEE 15th International Symposium on Consumer Electronics (ISCE)* (pp. 192–195). Available at https://doi.org/10.1109/ISCE.2011.5973811

Yan, D; Dan, Z. (2010). ZigBee-based Smart Home system design. In *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)* (Vol. 2, pp. V2–650–V2–653). Available at https://doi.org/10.1109/ICACTE.2010.5579732

Zamora-Izquierdo, M. A; Santa, J; Gomez-Skarmeta, A. F. (2010). An Integral and Networked Home Automation Solution for Indoor Ambient Intelligence. *IEEE Pervasive Computing*, *9*(4), 66–77. Available at https://doi.org/10.1109/MPRV.2010.20

Utilisation of DANGER and PAMP signals to detect a MANET Packet Storage Time Attack

Lincy Elizebeth Jim
RMIT University

Mark A Gregory
RMIT University

Abstract: The dynamic distributed topology of a Mobile Ad Hoc Network (MANET) provides a number of challenges associated with decentralised infrastructure where each node can act as the source, destination and relay for traffic. MANET's are a suitable solution for distributed regional, military and emergency networks. MANETs do not utilise fixed infrastructure except where connectivity to carrier networks is required and MANET nodes provide the transmission capability to receive, transmit and route traffic from a sender node to the destination node. In this paper, we present a Packet Storage Time (PST) routing attack where an attacking node modifies its storage time and thereby does not forward packets to the intended recipient nodes. In the Human Immune System, cells are able to distinguish between a range of issues including foreign body attacks as well as cellular senescence. This paper presents an approach using Artificial Immune System based Danger signal (DS) and Pathogen Associated Molecular Pattern (PAMP) signal to identify a PST routing attack.

Keywords: MANET; Packet Storage Time Attack; Artificial Immune System; Security, Danger Signal, Pathogen Associated Molecular Pattern

Introduction

A Mobile ad hoc Network (MANET) is formed by a group of mobile wireless nodes that do not require fixed infrastructure to maintain network connectivity (Giordano 2002). One of the many advantages of MANET is the absence of dedicated infrastructure to support packet forwarding and routing as each node acts as both host and router. MANET supports communications for military operations through to communications for the commercial sector including key roles in rescue or emergency scenarios (Loo, Mauri, & Ortiz, 2011). MANET can also be used to provide flexible education, health and business networks.

This autonomous nature of MANET makes it vulnerable to malicious active and passive attacks (Deng, Li & Agrawal 2002), and every node must be designed and built with the capability to respond to direct or indirect adversarial events. The mobility characteristics of

http://doi.org/10.18080/ajtde.v5n2.100

the ad hoc network facilitate independent management and control whilst part of one or more networks and this flexible design provides an opportunity for nodes to be compromised and affect the overall operation and efficiency of MANETs.

Based on the route discovery mechanism, the routing protocols in MANET can be classified into reactive or proactive (Mbarushimana & Shahrabi, 2007). Route discovery is initiated whenever there is a packet at a node in need of a route to the next node in the path to the destination node. Proactive routing protocols such as table-driven protocols are proactive where routes are computed beforehand and stored in the routing table so that routes will be readily available whenever any packet has to be transmitted.

In this paper, a detection system based on an Artificial Immune System (AIS) is proposed to detect ambushed or compromised nodes. The proposed framework utilises the principles of a dendritic cell algorithm which in turn mimics the Human Immune System (HIS), which has evolved into a sophisticated protector of the human body.

This paper is organised in the following five sections. Section 2 details key attack approaches found in existing in ad hoc networks and a brief overview of AIS. Section 3 provides the AIS based detection scheme. Sections 4 and 5 analyse the detection scheme and Section 6 contains the conclusion and work for future study.

MANET Attack Types

Security attacks on nodes in a MANET can be classified as either active or passive. Passive attacks involve snooping on the data exchanged in the network and often without the intention of altering the traffic. Passive attacks are very difficult to detect, because the network operation is not affected and they gather information about the network or pry on the communications between two or more nodes. This type of an attack may lead to an active attack if information gathered leads to an opportunity that would provide a positive outcome for the attacker. In passive attack a key facet that is the confidentiality of the network can be compromised.

In active attacks, the attacker alters the data being exchanged in the network thereby disrupting the normal functioning of the network and may launch an intrusive attack on a node (Deng & Agrawal 2002). The malicious behavior involves modification, injection or dropping packets and has a direct and immediate effect on network operation including affecting information security.

MANET nodes depend on battery power for operation and a loss of power efficiency may result from active attacks that lead to increased traffic (Perkins, 2008). Mobile nodes may

http://doi.org/10.18080/ajtde.v5n2.100

offer themselves as a relay node to forward data from other nodes in the network and providing a relay service can decrease the power available for the nodes use.

A brief description of the different types of attacks that occur in MANET is provided in the following sections.

Replay attack

In a MANET the topology changes permitting replay attacks where the attacker uses the strategy of storing control messages previously sent by a node (Adjih, Raffo & Muhlethaler, 2005; Goyal, Parmar & Rishi, 2011).

The attacker node resends the stored control messages which leads to genuine nodes updating their routing tables with stale information. This disturbs the normal operation of MANET.

Black hole attack

In a black hole attack the attacker node sends a false routing message claiming that it has the most conducive route to the destination, whereby leading all the genuine nodes to forward their packets to the attacker (Bala, Bansal & Singh, 2009; Mistry, Jinwala & Zaveri, 2010).

Flooding attack

In a flooding attack the attacker node sends multiple RREQ messages to a destination node that does not exist in the network (Bandyopadhyay, Vuppala & Choudhury, 2011; Yi, Dai, Zhang & Zhong, 2005).

As the destination node does not exist, none of the nodes will be able to send a Route Reply, leading to congestion and a network denial of service.

Wormhole attack

Wormhole attack is depicted in Figure 1. In this attack, the attacker nodes fabricate a route shorter than the original route, which in turn creates confusion amongst other nodes (Mahajan, Natu & Sethi, 2008). This attack is carried out by one or more nodes that create a tunnel between them whereby the attacker seizes packets and transmits to other nodes. The two colluding nodes send fake advertising messages that they have a single hop symmetric link between each other. These fake messages will be propagated to other nodes across the network thus compromising the shortest path routing calculations.

http://doi.org/10.18080/ajtde.v5n2.100

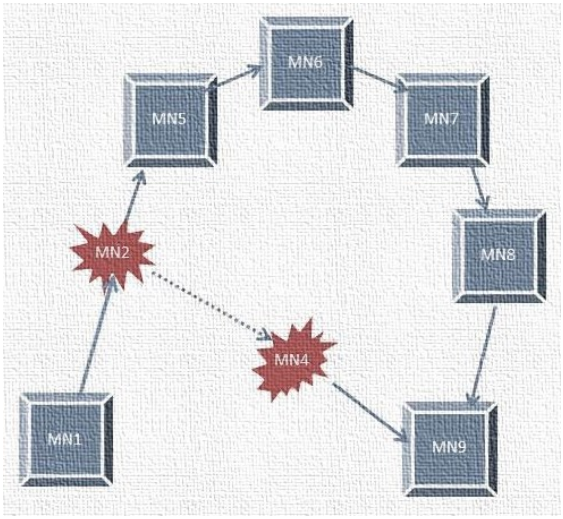


Figure 1 Colluding nodes in Wormhole Attack

The authors proposed a scenario where each node will keep track of the behaviour of its neighbours (Choi, Kim, Lee, & Jung, 2008). A RREQ message is sent by a node to the destination using its neighbour list. However, if the RREP is not received back within a stipulated time, the presence of a wormhole is identified and the route is added to the source node's wormhole list. Each node maintains a table which consists of RREQ sequence number and neighbour node identity. After sending a RREQ, the source node sets the Wormhole Prevention Timer (WPT) and waits until it overhears retransmission by the neighbour node. The maximum amount of time for a packet to travel a one-hop distance is WPT/2.

Artificial Immune Systems

AIS are intelligent and adaptive systems inspired by the human immune system toward real-world problem solving (Abdelhaq, Hassan & Alsaqour, 2011). AIS are adaptive intelligent systems "inspired by theoretical immunology and observed immune functions, principles and models, which are applied to complex problem domains" (Abdelhaq, Hassan, & Alsaqour, 2011).

A relatively recent immunological discovery known as Danger Theory now paves the way for designing more efficient, second generation AIS. The Dendritic Cell Algorithm (DCA) is a biologically inspired technique developed to detect intruders in computer networks (Abdelhaq, Hassan, Ismail, Alsaqour & Israf, 2011). The DCA is based on a metaphor of naturally occurring Dendritic cells (DCs), a type of cell which is native to the innate arm of the immune system (Abdelhaq, Hassan & Alsaqour, 2011). DCs are responsible for the initial detection of intruders, including bacteria and parasites, by responding to the damage caused by the invading entity. Natural DCs receive sensory input in the form of molecules, which can

http://doi.org/10.18080/ajtde.v5n2.100

indicate if the tissue is healthy, or in distress. These cells have the ability to combine the various signals from the tissue and to produce their own output signals. The output of DCs instructs the responder cells of the immune system to deal with the source of the potential damage. DCs are excellent candidate cells for abstraction to network security as they are the body's own intrusion detection agents.

The DCA is one of the most well-known Danger Theory contributions and utilises the role of the DCs in the HIS as forensic navigators and important anomaly detectors. DCs are defined as antigens presenting lymphocytes in the innate immunity; these lymphocytes play a key role in either stimulating or suppressing the adaptive immunity T-cells and hence controlling the immune system's response type.

DCA's capability as an anomaly detector algorithm inspires the use of a biological model to introduce a further DC inspired algorithm, which could detect other attack types in a MANET (Mazhar & Farooq, 2008; Abdelhaq, Hassan, Ismail & Israf, 2011). In addition, many of MANET's special characteristics and properties are similar to the innate immunity's abstract features; such as the openness and susceptibility of each to different types of attacks (Mazhar & Farooq, 2008).

AIS are increasingly being used to secure MANET because of its low communication and computational overhead (Gu, Greensmith & Aickelin, 2011). AIS based intrusion detection systems in MANET utilise the exemplar of discrimination between self and non-self. In this approach the system is first put in a learning phase, where the system learns the characteristics of a normal environment. Any changes that occur which do not match the normal environment are considered harmful. The problem with such an approach is limited fault detection. The behaviour of the normal environment during the learning phase cannot be taken as the prototype for trustworthy behaviour as a MANET environment keeps changing with time.

This approach becomes impractical once malicious nodes enter the network. As a result, identifying a valid route change due to mobility from a malicious node or advertisement of short routes by a malicious node can be challenging.

The properties of AIS (Mazhar & Farooq, 2008) such as being self-healing, self-defence and self-organising can be applied to meet the challenges of securing the MANET environment.

Proposed Attack-Packet Storage Time

Consider a MANET topology as shown in Figure 2, where the ad hoc network challenges to establish a route between the existing nodes is presented. In Figure 2, MN1 is the source

http://doi.org/10.18080/ajtde.v5n2.100

node and MN7 is the destination node. The intermediate nodes are MN2-MN6. In this scenario, consider the following available routes:

MN1-MN2-MN6-MN7

MN1-MN5-MN7

MN1-MN4-MN3-MN7

When a route is needed, the source node should take into account the battery power or energy of the participating nodes that provide a route reply.

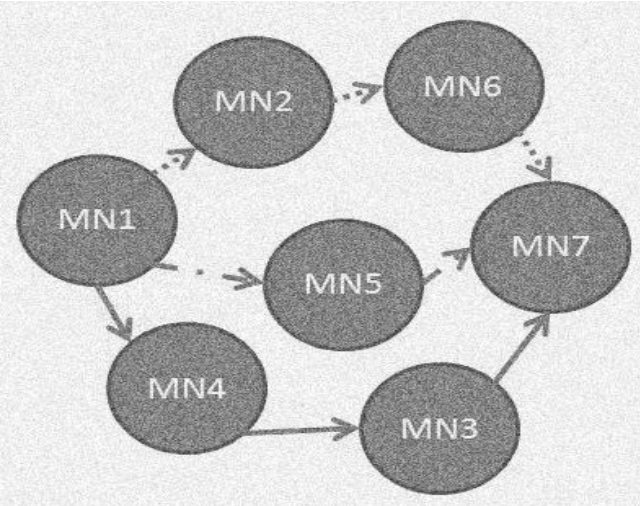


Figure 2 MANET scenario

In this scenario, a problem was identified where MN5 is an attacker/selfish node which does not want to expend energy to forward packets. This node is particularly interested in giving a RREP as it wants to maintain an updated routing table.

The Packet Storage Time (PST) attack is a novel concept introduced for the first time in MANET where each mobile node is incorporated with a buffer/queue. In this type of attack, the attacker modifies its own buffer/queue time to congest the network. When the packets are kept for a longer time than intended by each node, the packets become stale and the circulation of stale packets in the network leads to battery power wastage by the genuine nodes.

http://doi.org/10.18080/ajtde.v5n2.100

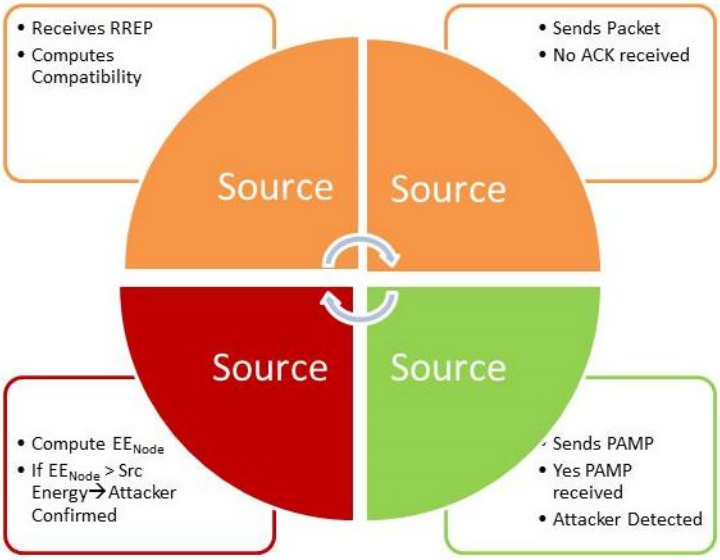


Figure 3 Proposed AIS algorithm

Proposed AIS Algorithm

A model based on the Danger Theory principle wherein each node is modelled as a DC is proposed. The presence or absence of danger is detected by the DC nodes, thereby identifying danger by indicating the presence of a malicious node. The DC nodes monitor the activity occurring in MANET to report any malicious node. The Pathogen Associated Molecular Pattern (PAMP) signal is utilised here to signify the presence of a malicious node in the network. Based on the concepts identified in the literature a mathematical model was constructed. The closer nodes should spend less energy to communicate. The battery life/energy of each DC node plays an important role while establishing routes. During route establishment, the energy of each node needs to be considered. Consider *n* to be an energy dependent variable. The energy associated with the source destination and intermediate nodes is assigned a weight which is dependent on the percentage of battery power that would be used during a route request and route reply communication. Based on above concept, the below given equation is formulated.

$$f(n) = \alpha n_s + \beta n_d + 2\gamma \bar{n} \quad (1)$$

where: $\alpha+\beta+\gamma=1$ and $\alpha, \beta, \gamma \in [0, 1]$ and $\bar{n}$ is the average of the energy among intermediate nodes, n_s is the source node energy, n_d is the destination node energy, α is the source node weight factor, β is the destination node weight factor, and γ is the intermediate nodes weight factor.

http://doi.org/10.18080/ajtde.v5n2.100

Consider the Effective Energy (EE) of Node *k*

$$\tilde{n}_k = F(n_k,h_k)=EE_{node(k)} = h_k * n_k \quad (2)$$

where h_k is the number of hops from node k to node s, and $F(n_k,h_k) \approx EE_{node(k)}$ should satisfy the postulates:

(1) If node *k* is far away from source node *s*, node *k* should have to take larger number of hops and more energy would be utilised which results in larger function value.
(2) If node *k* is closer to node *s*, node *k* should have to take lesser number of hops and lesser energy would be utilised which results in a smaller function value.

The effective mean energy of all the intermediate nodes is as follows

$$\tilde{n} = \frac{1}{m-2} \sum_{k=1,k\neq s,d}^{m} \tilde{n}_k \quad (3)$$

Combining (2) and (3) gives the Node Energy Momentous function

$$f(n)=\alpha n_s + \beta n_d + \frac{\gamma}{m-2} \sum_{k=1,k\neq s,d}^{m} h_k * n_k \quad (4)$$

From Eq (2) we get the Compatibility function

$$\hat{C}=1/F(n_k,h_k) \quad (5)$$

As Compatibility $\hat{C}$ increases the cost to establish the route between source and destination decreases which also implies that the node k has energy available for routing.

In a MANET, the source node initiates a route discovery whenever it should send a packet. The proposed Artificial Immune Systems Based Algorithm (AISBA) model consists of the following stages:

Normal

Consider the proposed AISBA model as shown in Figure 3. Initially the source initiates a route discovery in order to send a packet to a destination node and computes compatibility of the node from which it receives a RREP. The source node does not receive an acknowledgement (ACK) from the node that provided the RREP.

Attacker detection by using Danger Signal

When the source does not receive the ACK it activates its DC and sends a Danger Signal (DS); the good nodes acknowledge the Danger signal by sending Danger Signal received (DSrecvd). As the Danger Signal is not a priority signal there is no overwriting of the node buffer therefore the attacker node does not send back DSrecvd. This strategy is used when there is a smaller number of attacker nodes.

http://doi.org/10.18080/ajtde.v5n2.100

Attacker Detection

The source sends a high priority packet PAMP (high priority signal) message to the attacker node and the attacker node is forced to acknowledge receipt of the PAMP which indicates the presence of an attacker but this is not yet confirmed.

Attacker Confirmation

The source computes the node EE (EE_{node}) of the attacker node and compares the value with its own energy. If the EE_{node} happens to be greater than EE_{source} the presence of the attacker is confirmed.

The algorithm pseudo code is described as shown in Figure 4 and Algorithm 1. The source node broadcasts a RREQ and computes node compatibility for the nodes from which a RREP is received. The source node begins to send the packet and if an ACK is not received a high priority PAMP is sent. If the node is an attacker and it does respond with an ACK; this indicates the presence of the attacker node. The next step taken by the source is to compute the EE of the attacker, and a high EE value is used to confirm the presence of the attacker. This is also symbolically represented in the flowchart as shown in Figure 4.

Algorithm 1 Pseudo code of AISBA

1. Source Node broadcasts RREQ
   a) $Node_{src}$ broadcasts RREQ
   b) $Node_{intermed}$ sends RREP
2. Compute Compatiblity ($Node_{src}$,$Node_{intermed}$,Packet P)
   a) $Node_{src}$ computes compatibility of $Node_{intermed}$;
   b) **If** $Node_{src}$ sends packet and $Node_{src}$ does not receive Ack **then**
3. SendDS ($Node_{src}$, $Node_{intermed}$, DS_{send})
   a) $Node_{src}$ sends DS
   b) **If** $Nodei_{intermed}$ does not acknowledge DS **then**
   c) Detect if attacker or route error, do SendPAMP
4. SendPAMP ($Node_{src}$, $Node_{intermed}$,PAMPpacket PAMPp)
   a) $Node_{src}$ sends PAMP,
   b) **If** $Node_{intermed}$ acknowledges PAMP **then** attacker detected
5. Compute Effective energy of intermediate node ($EE_{nodeintermed}$)
   a) $EE_{nodeintermed}$ is computed by Source,

http://doi.org/10.18080/ajtde.v5n2.100

b) **If** $EE_{nodeintermed}$ >= $Node_{src}$, **then** presence of attacker is confirmed.

Simulation and Results

The ns-3.23 simulator was used to detect and confirm the presence of PST attacker using the AODV protocol. The simulation parameters used are shown in Table 1.

As can be seen in Figure 5 as the hop count increases in the network, the EE consumption by the mobile node is higher. As compatibility increases the cost to establish the route between the source and destination will decrease as can be seen in Figure 6. The route cost is a metric which is the ratio of transmitted control packets to the transmitted data packets. An increase in compatibility decreases the cost.

The average end-to-end delay (E2E) increases as shown in Figure 7 and during the PST as the attacker delays the packet whereas in an AISBA model the presence of an attacker is detected and confirmed thereby the packet will be forwarded via a routing path that does not contain the PST attacker. Hence the average E2E will be slightly higher for AISBA.

As can be seen in Figure 8, a PST attack causes packet loss to increase as the number of nodes increase, whereas with the AIS-based algorithm the packet loss is lower due to the proposed security improvement.

Table 1 Simulated Parameters

Simulator	Ns-3.23
Mobility Model	Random waypoint
Simulation Time	500s
Number of nodes	10-50
Traffic Type	UDP
Network Area	600m*600m
Mobility	6 m/s
Pause Time	5s
Transmission Range	50m

http://doi.org/10.18080/ajtde.v5n2.100

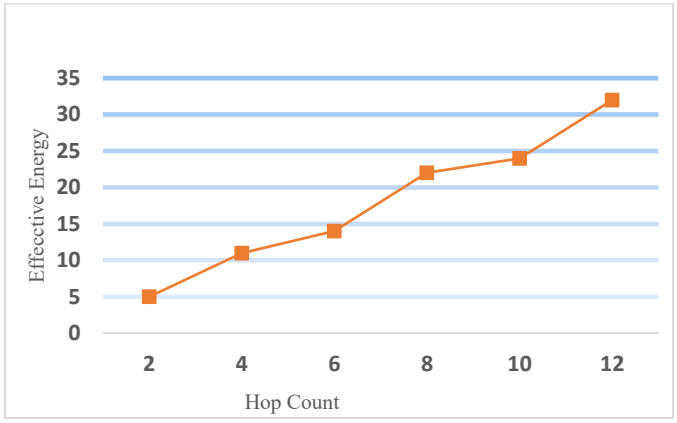


Figure 4 Effective Energy v/s Hop Count

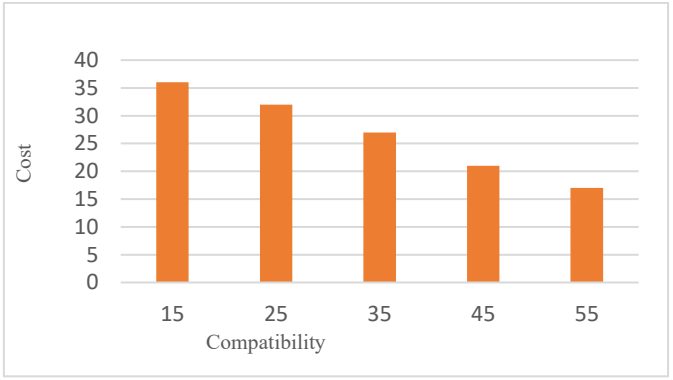


Figure 5 Compatibility v/s Cost

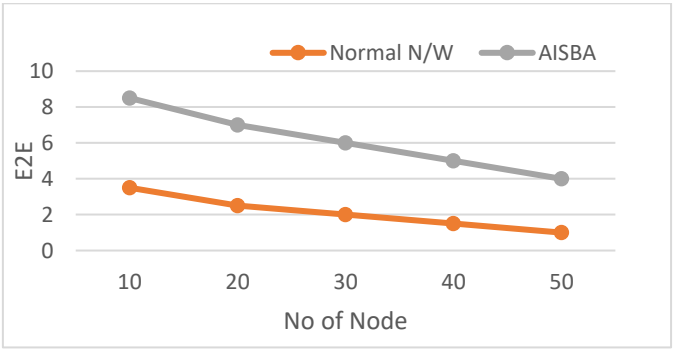


Figure 6 E2E v/s Number of nodes

http://doi.org/10.18080/ajtde.v5n2.100

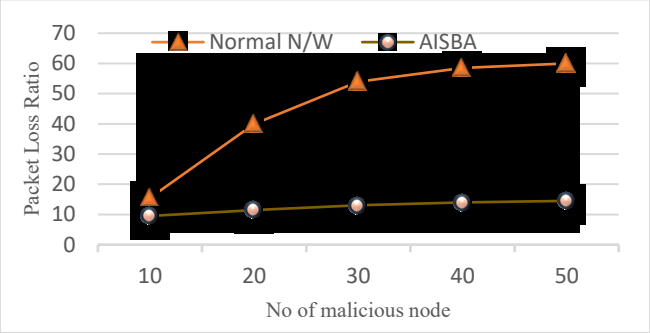


Figure 7 PST Attack-Packet loss v/s Number of nodes

Conclusion

In this paper, the PST attack is presented and the PST attack has been analysed using AIS principles and metrics including packet loss, delay and battery power. The solution proposed is robust and adopts AIS principles as an example of how AIS can be applied to MANET thereby reducing the effects of security events based on this attack type including inducing futile battery consumption. In future work, utilisation of a battery power metric with respect to the major nodes will be considered and will be included in the development of an improved security technique to increase MANET robustness.

References

Abdelhaq, M; Hassan, R; Alsaqour, R. (2011). *Using dendritic cell algorithm to detect the resource consumption attack over MANET*. Paper presented at the International Conference on Software Engineering and Computer Systems. https://link-springer-com.ezproxy.lib.rmit.edu.au/chapter/10.1007/978-3-642-22203-0_38

Abdelhaq, M; Hassan, R; Ismail, M; Alsaqour, R; Israf, D. (2011). Detecting sleep deprivation attack over manet using a danger theory-based algorithm. *International Journal of New Computer Architectures and their Applications (IJNCAA), 1*(3), 534-541. http://sdiwc.net/digital-library/detecting-sleep-deprivation-attack-over-manet-using-a-danger-theorybased-algorithm.html

Abdelhaq, M; Hassan, R; Ismail, M; Israf, D. (2011). Detecting resource consumption attack over MANET using an artificial immune algorithm. *Research Journal of Applied Sciences, Engineering and Technology, 3*(9), 1026-1033. http://www.airitilibrary.com/Publication/alDetailedMesh?docid=20407467-201109-201411110031-201411110031-1026-1033

Adjih, C; Raffo, D; Muhlethaler, P. (2005). *Attacks against OLSR: Distributed key management for security*. Paper presented at the 2005 OLSR Interop and Workshop.

http://doi.org/10.18080/ajtde.v5n2.100

https://www.researchgate.net/publication/242417041_Attacks_Against_OLSR_Distributed_Key_Management_for_Security

Bala, A; Bansal, M; Singh, J. (2009). *Performance analysis of MANET under blackhole attack*. Paper presented at the Networks and Communications, 2009. NETCOM'09. First International Conference on. http://ieeexplore.ieee.org.ezproxy.lib.rmit.edu.au/abstract/document/5384021/?reload=true

Bandyopadhyay, A; Vuppala, S; Choudhury, P. (2011). *A simulation analysis of flooding attack in MANET using NS-3*. Paper presented at the Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 2011. 2nd International Conference on. http://ieeexplore.ieee.org.ezproxy.lib.rmit.edu.au/abstract/document/5940916/

Choi, S; Kim, D.-y; Lee, D.-h; Jung, J.-i. (2008). *WAP: Wormhole attack prevention algorithm in mobile ad hoc networks*. Paper presented at the Sensor Networks, Ubiquitous and Trustworthy Computing, 2008. SUTC'08. IEEE International Conference on. http://ieeexplore.ieee.org.ezproxy.lib.rmit.edu.au/abstract/document/4545782/

Deng, H; Li, W; Agrawal, D. P. (2002). Routing security in wireless ad hoc networks. *IEEE Communications magazine, 40*(10), 70-75. http://ieeexplore.ieee.org.ezproxy.lib.rmit.edu.au/abstract/document/1039859/

Giordano, S. (2002). Mobile ad hoc networks. *Handbook of wireless networks and mobile computing*,325-346 http://au.wiley.com/WileyCDA/WileyTitle/productCd-0471419028.html

Goyal, P; Parmar, V; Rishi, R. (2011). Manet: vulnerabilities, challenges, attacks, application. *IJCEM International Journal of Computational Engineering & Management, 11* (2011), 32-37. http://skirubame.ucoz.com/_ld/0/29_Topic_1-MANET_v.pdf

Gu, F; Greensmith, J; Aickelin, U. (2011). The dendritic cell algorithm for intrusion detection. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2824971

Loo, J. H; Mauri, J. L; Ortiz, J. (2011). *Mobile ad hoc networks: current status and future trends*: CRC Press. http://www.crcnetbase.com.ezproxy.lib.rmit.edu.au/doi/pdf/10.1201/b11447-1

Mahajan, V; Natu, M; Sethi, A. (2008). *Analysis of wormhole intrusion attacks in MANETS*. Paper presented at the Military Communications Conference, 2008. MILCOM 2008. IEEE. http://ieeexplore.ieee.org.ezproxy.lib.rmit.edu.au/abstract/document/4753176/?reload=true

Mazhar, N; Farooq, M. (2008). A sense of danger: dendritic cells inspired artificial immune systems for MANET security. *Proceedings of the 10th annual conference on Genetic and*

evolutionary computation. *Atlanta*, GA, USA, ACM: 63-70.
http://dl.acm.org.ezproxy.lib.rmit.edu.au/citation.cfm?id=1389105

Mbarushimana, C; Shahrabi, A. (2007). *Comparative study of reactive and proactive routing protocols performance in mobile ad hoc networks*. Paper presented at the Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on
http://ieeexplore.ieee.org.ezproxy.lib.rmit.edu.au/abstract/document/4224182/

Mistry, N; Jinwala, D. C; Zaveri, M. (2010). *Improving AODV protocol against blackhole attacks*. Paper presented at the Proceedings of the International Multi Conference of Engineers and Computer Scientists.
http://www.iaeng.org/publication/IMECS2010/IMECS2010_pp1034-1039.pdf

Perkins, C. E. (2008). Ad Hoc Networking 2001. *Boston, Adison-Wesley*.
https://www.pearsonhighered.com/program/Perkins-Ad-Hoc-Networking-paperback/PGM75272.html

Yi, P; Dai, Z; Zhang, S; Zhong, Y. (2005). A new routing attack in mobile ad hoc networks. *International Journal of Information Technology, 11*(2), 83-94.
http://www.intjit.org/cms/journal/volume/11/2/112_6.pdf

http://doi.org/10.18080/ajtde.v5n2.100

Fact or Fraud?

the epidemic which struck Telecom Australia from 1983 to 1986

Ian Campbell

Telecommunications Association

Abstract: Epidemic - a widespread occurrence of an infectious disease in a community at a particular time - outbreak, scourge, plague.

About 1976 concerns emerged in Australia about the potential for new technologies to seriously reduce employment. The debate reached a peak in 1978 when industrial action taken by the Australian Telecommunications Employees Association (ATEA) threatened to shut down the Australian telecommunications network.

From 1975 growth in calls to the directory assistance service rocketed as did the related operating costs. To maintain the quality of customer service and contain operating costs, Telecom began to deploy a nation-wide computer- based directory assistance system (DAS/C) from 1982. In 1983 an unexpected medical syndrome arose in one of Telecom's directory assistance centres.

Over the next three years, the syndrome rapidly spread through other directory assistance centres, other areas of Telecom and some areas of the public service. The media, academia, legal practitioners and others attracted to the "problem" generally accepted the union view that the DAS/C system was the cause and the syndrome was labelled Repetitive Strain Injury (RSI). Medical costs and compensation claims mounted reaching $130 million in 1989.

While RSI has become a well-known syndrome over the last 40 years, no outbreaks of the extent and severity experienced in Telecom appear to have been recorded over that period anywhere in the world. There still appears to be little scientific evidence of the link between the injury of the reported scale and the workplace.

This is the story of the rise of the RSI phenomena in Telecom over the period 1983-86.

Keywords: Telecommunications, Telecom Australia, Directory assistance, RSI.

http://doi.org/10.18080/ajtde.v5n2.107

Introduction

The Australian Telecommunications Commission (Telecom) was established in 1975 as a government owned business with open-ended regulated monopolies; building and operating the national telecommunications network and the sale, rental and maintenance of certain customer premises equipment. These monopolies allowed Telecom to maintain a share of the Australian telecommunications market approaching 90%.

The original intention was that the new Telecom operate on "commercial business principles", with personnel and other employment policies, including industrial relations, suited to the new business and independent of the Public Service Board. In fact, Telecom was implemented with minimal commercial experience, and the personnel and other employment policies, processes, and culture of the public service were retained.

The combination of the open-ended monopolies, the public sector legacy and policies, the public service culture, and the public service unions, unrestrained by competition, was to prove a major obstacle to innovation and progress in Telecom. Change was introduced within these limitations and within the management's preference for change on its own terms with minimal risk and, during the 1980's, within the tolerance of the Labor Government.

This background is crucial to understanding the sudden, unexpected, intense and prolonged appearance of RSI in Telecom and its far lower, sporadic appearances elsewhere.

Disclosure

This paper provides a reflective historical paper about the RSI occupational health and safety issue and a case study of the effects of concerns about RSI in Telecom in the first half of the 1980's.

The paper is supported by a number of records of the period, including business plans, business cases, and trading statements, as listed under "References". The records are incomplete but are sufficient to support the points made. A number of these records no longer exist or are not easily accessible such as those in the archives of the Australian Telecommunications Commission and Telstra.

The author has no qualifications, training or experience as a physician or in the fields of medicine, physiotherapy, or occupational health and safety. Opinions and judgments are within this context and are the author's unless otherwise stated. Those of Telecom are expressed using standard private sector criteria including growth, market share, customer service and profit, rather than using public service criteria.

http://doi.org/10.18080/ajtde.v5n2.107

The Directory Assistance Service

People needing a local telephone number to make a call could obtain the number from the local published directory or, if a local directory was not available, from the directory assistance 013 service. People needing an out-of-area number called the directory assistance. 0175/0171 service. Directory assistance employed operators, mostly women, to provide these two number services.

During the year, in a telephone service area, new services were connected, some existing services cancelled, and some services changed, each generating a change in the directory listing information - name and address etc - applying to the numbers for those services. By the time the published directory for the next year was delivered, the previous directory might be 20-30% in "error"; wrong numbers or missing numbers, and this factor reinforced the need for the directory assistance service.

From 1976 Telecom made a major investment to improve the attractiveness, accuracy and availability of the published directories to encourage usage and to attempt to contain the growth in calls to directory assistance. In 1977 studies of calls made to directory assistance reported that, of all calls for numbers in the local area, about 70% of people asked for numbers correctly printed in the current directory; five years later the proportion was roughly the same. That is, either the customer did not have a copy of the local directory - less likely as distribution greatly improved - or the customer was too lazy to look up the directory. 80% of calls to directory assistance were for business numbers implying that some businesses were using the service as a reference for their operations; for example, verifying telephone numbers and addresses for credit applicants.

The Directory Assistance Process to 1981

In 1976 the directory assistance service was provided by 1650 operators working in about 100 call centres located around the country, with the largest centres in the capital cities.

The following is broadly indicative of the call handling process.

When a customer called for a number, the operator asked for the name and address and an indication (if known) of the rough time when the service was connected. If the number had been connected before the issue of the last directory, the operator consulted a copy of that directory. If the number had been connected more than a month ago, the operator consulted a printed "monthly update" for the appropriate month. A number connected within the last month would be obtained from a printed "weekly update".

http://doi.org/10.18080/ajtde.v5n2.107

The “monthly” and “weekly” updates were provided by Telecom's directory publishing business through the directory printers. The delay between when a telephone was connected in the field to receipt of the number on a printed update in the directory assistance centre was typically 5-10 days.

The sight of a large metropolitan directory assistance centre was astonishing; apart from the air conditioning and modern lighting, the work process looked a throwback to a 19th Century workshop. Piles of paper directories, paper monthly updates and paper weekly updates were located around operators who could take up to a minute and a half to find and provide the requested number. Some requests for numbers could not be satisfied.

Measurement of the performance of the directory assistance centres was unacceptable for such a poor service and costly operation. There were perhaps six main performance parameters for assessing the quality and efficiency of the service;

- the **number of calls entering the queue** at each directory assistance centre,
- the **number of calls answered**,
- the **speed or answer**; the elapsed time between when the call entered the queue and when the call was answered,
- the **average operator work time**; the elapsed time between connection of the call and disconnection of the call after the requested number was provided,
- the **number of calls for which the requested number was not able to be provided,**
- the **number of calls handled per operator per year**.

The striking fact was that, of the six parameters, only one was measured routinely (calls answered), three were measured by sporadic and doubtful sampling or estimates **-** speed of answer, average operator work time, number of calls handled per operator per year - and two were not measured at all. The union routinely refused to allow assessment of an individual operator’s performance, even to identify the need for further training.

Tables 1 and 2 provide some indication of performance. Note that the data is incomplete and was not available for some years from some centres; in 1980/81 data for three of the six states was not collected, including Sydney, the largest centre.

The Status of the Directory Assistance Service in 1981

The future performance of the directory assistance service looked bleak. The 22 million calls to the service in 1976 had increased to 78 million in 1981. The directory assistance service

http://doi.org/10.18080/ajtde.v5n2.107

was very costly to provide and the costs were rocketing; operating costs were $4.3 million in 1971, $17.7 million in 1976, and $38 million in 1981, and were projected to reach $63 million in 1985. Costs were being driven by the rapid growth in calls, an obsolete manual process, the high labour intensity of the service, monopoly wages and conditions, union obstruction to productivity improvements, and union demands that operator numbers be increased to handle the ever-growing calls (rather than improve productivity).

Table 1: Directory Assistance - Speed of Answer for calls Connected to Metropolitan 013 Centres - 1977/81

Year ending 30th June	1977	1978	1979	1980	1981
	% of calls answered within 10 seconds				
Sydney	76.3	79.9	80.0	81.0	
Melbourne	-	72.1	79.1	74.7	77.9
Brisbane		76.4	82.6	80.0	71.0
Adelaide	-	80.2	82.4	74.9	82.3
Perth	80.7	77.0	79.3	78.0	
Hobart	75.5	76.4	85.0	88.4	

Note 1. The standard for the speed of answer was 90% of calls to be answered within 10 seconds.
Note 2. Source: Network Operations Branch, Engineering Department, Headquarters.

Table 2: Directory Assistance - Average Operator Work Times - Metropolitan 013 Centres - 1977/81

Year ending 30th June			1977	1978	1979	1980	1981
					seconds		
Sydney		-	49	48.7	55.0	50	
Melbourne			-	-	-	-	50
Brisbane			-	83	-	65	69
Adelaide			-	54	-	50	50
Perth		-	52	52	47	46	
Hobart	-	-	-	-	68	50	

Note 1. The average operator work time was the number of seconds between connection of the call to the operator and disconnection of the call after the requested number had been provided. The time for a caller to obtain a number was the waiting time in the queue plus the average operator work time. Note that a significant number of callers hung up before connection to the operator as they did not have time to wait for service, and these calls were not measured.
Note 2. Source: Network Operations Branch, Engineering Department, Headquarters.

The wages, conditions, working practices and working rules for the directory assistance service were significantly out of line with private sector standards, further inflating labour costs - see later.

http://doi.org/10.18080/ajtde.v5n2.107

Hindered by the union, the management approach was to attempt to reduce the growth in the number of operators which, as calls continued to climb, resulted in a degraded standard of service - longer wait times to answer the calls and more calls not answered. In 1981 the service quality and almost every other aspect of the operation was approaching unacceptable; among the stakeholders - Telecom, customers, management, staff, and the unions - the interests of the operators and the operators' union appeared to be the primary concern and the customers ranked last.

Unless aggressive action was taken to improve productivity, the only way to avoid a substantial rise in the number of operators and a higher rise in operating costs was to markedly improve productivity or further and significantly degrade the quality of service.

A Directory Assistance Computer-based System (the DAS/C System) from 1982

By 1981 advanced computer based switching systems were available for aggregating calls from around a region and distributing the calls to directory assistance centres according to rules set for managing the queue at each centre. Such systems coupled with far fewer call centres offered a far more efficient aggregation, distribution and queuing of calls that could provide an acceptable quality of service with 10-25% fewer operators with some margin for call growth. Reducing the number of centres and operators was rejected by union.

In 1977 a Telecom manager visited the two leading telecommunications businesses in the USA, AT&T and GTE. Their directory assistance operations were far superior to Telecom's and their plans would transform the service over the next five years. The operators still used paper records but these were far better organised and provided a higher level of performance; for example, the speed of answer was usually 90% within 10 seconds and average working times of less than 50 seconds were common.

Central to their plans was a computer-based system being trialled which would revolutionise the service and produce a step function improvement in customer service and efficiency. **AT&T estimated that each second of average work time saved by AT&T in directory assistance was worth about $1 billion per year in cost savings.**

In 1979 a trial of the US computer system began in the Sydney 013 centre which quickly demonstrated the worth of the system. After calling for public tenders and against union opposition, a nationwide system from IBM was progressively installed in directory assistance centres around Australia from 1982 and was fully deployed in 1986.

The national system consisted of **three sub-systems**. A **national directory assistance data base** was supplied from the White Pages data base consisting of listings collected from

http://doi.org/10.18080/ajtde.v5n2.107

around Australia which were compiled and edited by the directory publishing units. Every day the national directory assistance data base was updated from the White Pages data base. A **national call management system** collected calls to directory assistance from around Australia and dynamically and optimally distributed the calls to the directory assistance centres, depending on the call queue at each centre, and measured and recorded call data. At the directory assistance centres operators at **DAS/C computer terminals** answered the calls and interrogated the data base to find the numbers requested, and the system measured the inquiry and call handling characteristics and times.

The national system offered an enormous improvement in customer service and productivity, substantially lower operating costs, and for the operators, an easier work process, a more pleasant workplace and far higher work satisfaction. For the first time

a complete data base for the service was available including such as calls answered, calls completed, operator handling times and other operator statistics.

Computer generated voice announcements and related technology could further reduce the time for an operator to handle a call. For example, a recorded voice announcement at the start of each call could ask to "please state the name and area" related to the requested number, and "please have a pencil ready to record the number". At the end of the call a recorded voice sign-off such as "the requested number is " (with an option to repeat) provided the number to the caller when found (releasing the operator for the next call). In several years computer voice recognition technology could further reduce call handling times. Computer voice announcements and voice recognition were expected to be initially unpopular with customers and would generate complaints to a regulator. However, properly designed and introduced over time the announcements would likely become accepted, particularly as the published directories and the directory assistance service improved.

By 1986 the IBM DAS/C system operated in 55 directory assistance centres around Australia with 617 operator positions. The elapsed time between when a Telecom sales office recorded an order for a new, changed or cancelled telephone service and when the directory details were recorded on the DAS/C data base - to be available to directory assistance positions throughout Australia - was typically 24 hours; this compared to the old time of up to 10 days. The DAS/C data base was available to every operator position and was one of the largest in Australia.

The new DAS/C system was needed urgently. As previously mentioned, in 1982, the first year of the program, the directory assistance service was a struggling mess. Service quality and operator productivity were falling, and the number of operators had risen from 1650 in 1976

http://doi.org/10.18080/ajtde.v5n2.107

to 2300 (an increase of 39%). When installation had been completed in 1986 calls to directory assistance had climbed from 78 million in 1981 to over 145 million.

The average speed of answer of calls rose from 38-50% within 10 seconds to 59-86% within 10 seconds, and the average operator work times to complete a call fell from 51-78 seconds to 39-42 seconds.

However, this performance was still woefully short of the level others were achieving with a similar system, particularly in the USA.

The whole project was strongly opposed by the union at every stage.

The Industrial Relations Context

It is important to understand the technical and industrial climate within which the DAS/C system was planned and deployed.

Broadly, from 1975 until 1985 there was a revolution in the application of digital technology and software to telecommunications networks and computing. Digitisation offered a wide range of new telecommunications services, a higher revenue potential, a far higher standard of network performance and reliability, lower equipment and construction costs, lower operating costs, higher capital and labour productivity, and lower accommodation costs. The benefits of the new technologies prompted Telecom to launch three major projects towards the late 1970's.

The first project, automation of telephone exchanges and digitisation of the links between from 1978, threatened enormous changes in the technical and operator workforces. Exchange automation offered automatic connection of most long distance and international calls, eliminating most of the telephone operators in the Manual Assistance Centres (MAC) who manually connected those calls in the past.

The second project, the DAS/C system and future computer enhancements - the trial from 1979 - was facilitated by network automation, would completely change the skills required of the operators and could potentially reduce the number of operators by up to 30%.

The third project, was the District Customer Record Information System (DCRIS), a mainstream computer system launched in 1976 which processed telephone service orders and converted more than three million paper records to a computer data base. DCRIS completely restructured the work process and workforce in over 80 districts, changing the skills required and potentially significantly reducing staff.

http://doi.org/10.18080/ajtde.v5n2.107

These three projects and others would create huge changes in the network, service delivery, and the technical, operator and clerical workforces. There would be changes in the workforce structures, new skills would be required, obsolete skills would be discarded, some jobs would be de-skilled, and there would be large reductions in people employed. If the full potential of the technology was applied, unrestrained by industrial factors, over the 1980's the reduction in the workforce might exceed 10,000 and a similar number of jobs might have the skill mix changed, many radically. With a Labor Government in power for most of the 1980's, Telecom's network monopoly imposing no downside on industrial action, and the militancy of key Telecom unions, there was no prospect that this would occur.

The union covering the technicians working in the network was the Australian Telecommunications Employees Association (ATEA), a very strong union with a high membership which became highly militant after 1975.

Telecom announced in 1978 a $2 billion plan to computerise its exchange system. The plan would significantly reduce the number of technicians employed; some jobs would be up-skilled, many would be downgraded, a large number would be eliminated, and career opportunities would be seriously reduced or extinguished. For example, exchange automation eliminated the need for a maintenance crew at most exchanges - as many as 14 in an exchange crew.

The announcement triggered a blunt reaction from the ATEA. The union banned repairs and maintenance in a number of areas. Within one month the telecommunications networks in NSW and Western Australia were near collapse, and Telecom was losing $1 million a day in revenue; a few days later the network failures had spread to Victoria and South Australia.

The settlement was a triumph for the ATEA; a moderation of the workforce changes and reductions, some restructuring of the maintenance centres to preserve some job satisfaction, more career opportunities and job classifications, and back pay for about 70% of the workers who were stood down. The union also won a measure of control over new technology with concessions about prior consultation and involvement in trials of new technology.

The union covering the telephone operators, including the directory assistance operators, was the Australian Telephonists and Phonogram Officers Association (ATPOA).

Exchange automation allowed customers to dial most long distance and international calls automatically rather than through an operator, making redundant most of the operators who manually connected those calls. Telephonists in metropolitan areas fell as the demand fell for operator connected calls when a manual exchange were replaced by a computer controlled exchange. Telephonists in each country town vanished overnight when the

http://doi.org/10.18080/ajtde.v5n2.107

manual telephone exchange was replaced. Telex operator numbers were declining as the service became obsolete. The trial of the DAS/C system in Sydney 013 centre and its deployment nationally from 1982 would potentially reduce the directory assistance workforce by 30%, change the skills of all the remaining operators and relocate as many as 50% of those operators.

Noting the example and success of the ATEA, the ATPOA had also become strongly militant. The ten-year period after 1975 saw aggressive, bitter and often prolonged confrontations between the union and management. By the early 1980's few, if any, white collar unions

(counting the ATEA as "blue collar") in Australia had encroached into the managerial prerogative as far as the ATPOA.

The experience of Telecom's middle management during the 1978 technicians' industrial action indicated that support for middle management from top management (and the ruling Labor government) during any industrial action by operators was unlikely.

There were perhaps five main demands the ATPOA made regarding the MACs and DAS/C plans; the number of operators employed be maintained, wage increases, improved redundancy benefits for those who lost their jobs, working arrangements that improved output would be opposed, and acceptance of the "Green Book". The Green Book, agreed in 1984, was a detailed document listing Telecom's obligations and commitments to telephone operators regarding employment policies, manual assistance management, and recruitment and redeployment which included most of the ATPOA's demands. As a statement of principles, it was a breakthrough for the ATPOA similar to the technological change protocols agreed with the ATEA in 1978.

In the short term the Green Book maintained the MACs at their stated locations and a number of other operating arrangements until the end of the 1986/87. One clause proposed that new products be assessed which were intended to maintain the level of operator employment. One new product proposed by the union was a 24-hour Community Information Service providing information and referrals on a wide range of subjects such as bus timetables in Melbourne's western suburbs, child care facilities in Fremantle, and Chinese painting classes in Townsville. The union's concept was of "the socially aware, helpful, local telephonist always at your service". The service was to be free and financed by Telecom's customers.

Over the five years to 1985 the union mounted a strong public relations and industrial campaign aimed at hindering the MAC and DAS/C plans and maintaining employment. Typical of the public campaigns was the Victorian members' use of the media and local events to build community and political support against the closure of manual exchanges in

http://doi.org/10.18080/ajtde.v5n2.107

Hamilton, Horsham, Swan Hill, Dromana and Frankston which would eliminate around 120 operators. One result was that Telecom was forced to slow deployment of the MAC plan by delaying the closure of many exchanges.

With the MAC plan proceeding, albeit more slowly, the union became more inventive and aggressive with actions to slow or prevent the progress of the DAS/C system. Claims were made for an increase of 20% in wages justified as a share of the productivity improvement, reduced hours of work (from 34 hours per week) and an alleged "shriek" phenomenon in 1980.

Table 3: Telecom - Exchanges & Telephone Operators - 1982/88

Year		82/83	83/84	84/85	85/86	86/87	87/88
Exchanges	Metro Auto	531	541	535	586	586	N/A
	Country Auto	4087	4131	4204	4216	4333	N/A
	Country Manual	735	622	488	360	237	N/A
Telephonists		7518	7038	7070	7032	6315	6220

Source: “Resistance on the Line” - the author has not validated the data.

The most damaging tactic was an alleged range of workplace injuries emerging in 1983 supposedly caused by working with the DAS/C keyboard; with the union's support the new condition was labelled Repetitive Strain Injury (RSI). RSI is covered in the next section.

The shriek was a high-pitched spike of electronic noise transmitted without warning through the network which was intensified in an operator’s ear by the close-fitting, insulated headset. Telecom’s 1981/82 Annual Report reported that there had been 174 reported cases of "acoustic trauma" while the union claimed that the real figure was over 4,000.

Despite the delaying tactics, from 1984 to 1987 the number of manual exchanges in rural areas dropped from 622 to 237, the number of telephone and directory assistance operators was in freefall, and the diminishing number of members threatened the survival of the union. By 1988 there were around 17% fewer telephonists than in 1983, with country closures the single largest cause of the losses, but full exploitation of the DAS/C system and voice messages and recognition could reduce operators by another 10-20%. In 1988 the ATPOA’s membership fell below a viable level and the union was merged into a branch of the ATEA, the technicians' union.

Repetitive Strain Injury - The Problem of Definition

Most definitions of RSI link symptoms such as pain with repetitive tasks. For example:

http://doi.org/10.18080/ajtde.v5n2.107

"RSI is a condition where pain and other symptoms occur in an area of the body which has done repetitive tasks (often the arms or hands)." And "Repetitive strain means strain related to actions which are frequently repeated."

RSI was used in the 1980s to describe the Telecom epidemic. It is still used for the condition in the UK and parts of Europe. Australia decided to adopt “occupational overuse syndrome” in the late 1990s (rsi.org.au).

A more probing observation (Ewan, Lowy and Reid, 1991) was that "RSI (is) a non-specific and controversial constellation of work-related hand, arm and neck symptoms (which) became endemic in Australian industry in the early 1980s".

A similar observation was that RSI is a poorly named condition usually applied to people with non-specific upper limb pain in occupational settings. In 1986 the Royal Australasian College of Physicians strongly discouraged the use of the term RSI as the "description and case definition of the syndrome without implying causality is a prerequisite for adequate clinical investigation into the pathophysiology of the disorder". In other words, the name contravenes the basic principles of taxonomy by assuming or implying findings and causality that have not been established.

The RSI Epidemic in Telecom

In 1983, within the first full year of implementation of the DAS/C system, some operators reported non-specific pain in the such as the fingers, hands, wrists, forearms, neck or shoulders.

In 1984 the union announced that there was a link between the repetitive nature of the keyboard activity and the alleged "injuries". With the following publicity, a number of names emerged from the medical and occupational health professionals and academics to describe this new medical disorder, including occupational overuse syndrome, occupational cervicobrachial disorder and the like. All of these terms and the label RSI enshrined the union's claim of a conclusive link between the repetitive physical activity and the reported range of injuries.

By February 1985 the union claimed that the number of reported RSI cases had risen to around 1,200, an increase of 630% in an eight-month period, due to the increasing usage of keyboards coinciding with cutovers to the DAS/C centres for the directory assistance service and cutovers from manual to automatic exchanges for the telephone call connection service.

Dr. Bruce Hocking (1989), Telecom's Director of Occupational Health, reported the following statistics in the Medical Journal of Australia in 1989.

http://doi.org/10.18080/ajtde.v5n2.107

In Telecom **between 1982 and 1985 there were 3,976 reports of "RSI".** The occupation most affected were telephonist operators with 1,886 complainants, a rate of 343 per 1,000 keyboard staff members over five years. Other Telecom occupations affected were clerical workers (1,421, rate 284 per 1,000), telegraphists (17, rate 34 per 1,000), and process workers (235, rate 116 per 1,000). Women accounted for 3330 (83%) of all reports. In the telephonist group, 27% of female and 20% of male staff members were affected; for women, younger staff members were more affected. 644 staff claimed to be affected for more than 26 weeks. The costs of the treatments, rehabilitation, compensation and redundancies exceeded $15 million, including $1.8 million in medical costs.

It is interesting that, as the peak approached in 1984, union records indicate that Hocking commissioned an occupational physician, Dr Colin Mills, to study and report on the incidence of RSI in Perth. Mills reported that **some 20% of the operators surveyed had a "diagnosed repetitive injury" and a further 24% displayed signs of an emerging problem".** "Every day they continue to work in the same environment using the same techniques they risk progression of their injury". Mills clearly had concluded that there was a link between the injury and the work.

Note that in an average directory assistance operator workforce of about 2,000 over four years some 1,886 operators complained of an injury, an extraordinary number; 34% of operators claimed to be stricken.

Reactions of the Stakeholders

Telecom's epidemic drew the media and professionals like flies to ordure.

Urged on by the union, supported by an uncritical, often sensationalist media, a new occupational health and safety industry emerged which was encouraged by certain elements of the medical profession, academia and personal injury lawyers, all focussed on Telecom's directory assistance centres.

At the height of the epidemic, a press agency was able to supply 40 to 50 clippings on the subject each month from major newspapers and periodicals, and there were articles in medical journals and other periodicals of all kinds. There were discussions about RSI on talk-back radio and documentaries on television, and numerous workshops, seminars, lectures, government papers, and several books on the subject. The issues were irresistible with all the features for successful reporting; widespread injury, incompetent or hostile doctors, controversy, unions, large awards of money to some workers, injustice, high emotions, new technology, and experts who disagreed. A woman with RSI was reported to have committed suicide by hanging in Adelaide, South Australia; it was the view of one

http://doi.org/10.18080/ajtde.v5n2.107

medical expert in the subject that, contrary to the headline, it is likely she suffered from undiagnosed or improperly treated depression.

During this time, ergonomics consultants flourished in Australia. Remedial workplace actions were advocated such as ergonomic redesign of the workstation, keyboard layout, arm use and sitting posture, regular strengthening exercises, and regular rest breaks. There was an avalanche of advice with experts from various countries. For example, people claiming seating expertise gave different opinions; Coe of New Zealand advocated low seats which conflicted with conventional ergonomic advice from Cakir of Germany, while Mandal of Denmark advocated a forward tilted seat, and Grandjean of Switzerland advocated a more reclining back-support.

The most-often prescribed treatments for early-stage RSI included drug therapies such as anti-inflammatory medications combined with passive forms of physical therapy such as rest, splinting, massage, physiotherapy and similar.

This was followed by the rise of firms specializing in "rehabilitation" supported by government policy. One such business company grew from two physiotherapists to over 160 staff in almost a dozen regional offices around the country, only to suffer investigation by government agencies for irregular claims for payments, and then collapse seven years after starting.

Sociologists and others identified a number of models or stereotypes in their analyses. For example, the "Noble Worker" whose complaint was believed to be genuine; the "Migrant Arm" of a migrant worker thought to be a malingerer, similar to the bad back syndrome; the "Kangaroo Paw", for workers who reported a unique Australian (non) disease, a diagnosis of RSI but with no evidence of a disease process; and the "Featherbedded Worker" who was encouraged by unions to expect ever easier work conditions such as lower key-stroke rates, lengthy rest breaks, daily exercise programs and the like.

Doctors too were stereotyped such as "The Caring Doctor" who leaned towards comfort and support rather than objective diagnosis and treatment; the "Jeremiah Doctor" who provided bleak diagnoses with insufficient empathy; and the "Insurance Doctor" who was considered biased towards the principal.

There was a vigorous debate among the academics and health professionals about the causal factor in the rise of RSI in Australia with a number nominated as important or possibly important. These included work and work characteristics such as changing work organization, worker factors such as migrant workers, and white middle-class female workers. Other possible causes included Australian trade unions, laws on workers' compensation, the medical profession, the legal profession, and the media.

http://doi.org/10.18080/ajtde.v5n2.107

One view was that this was a genuine condition, recognised by many medical professionals and organisations around the world, including the UK's NHS and the World Health Organisation; RSI was a medical condition that could be diagnosed, and biomechanical factors must be rectified to prevent further injury. Others argued that psychosocial factors within and outside the workplace caused a neurosis that was not occupational in origin and not liable for compensation. A third group argued that RSI was a hoax consciously used by workers and encouraged by the union to gain concessions including compensation, sick leave payments and more generous redundancy conditions and payments. A fourth theory was that the RSI afflicted people who are, in essence, healthy but experience pain or were encouraged to become patients with pain, redefining a state of health to one of illness. Yet another view was that, as no link could be found, inaccurate presentation by the media of the condition and the causes could increase the prevalence of RSI; some research reported that "failure to provide a balanced scientific analysis of RSI can stigmatize individuals" and can create contagion in colleagues.

A more scathing view was that for claims of RSI the degree of repetition was never quantified and injury was simply assumed because the claimant was a worker; some claimants alleging RSI did work which had little or no repetitive component and strain was never demonstrated. At the time of these claims, workers doing similar tasks in the same or comparable businesses did not report symptoms. In most cases, if not all, no physical injury was ever demonstrated by such as through organ imaging.

Two strongly different views emerged from the debate:

- physical injury occurred in practically all cases. The injury was in muscle and due to some unspecified pathology caused by overuse or could be confirmed by electron microscope studies, or came from the neck and could be demonstrated by a nerve stretching test, or was due to excess muscle tension.
- there was no evidence of injury in these cases, and complaints of pain and other symptoms were due to psychological causes. Such views were put forward by the Royal Australian College of Physicians, the Australian Hand Club (Australian surgeons specializing in surgery of the hand and forearm), orthopaedic surgeons, rheumatologists, psychiatrists, and others.

Supporters of each theory appeared to ignore the others. Physicians supporting the second view were regarded by the first group as either being ignorant, in the pay of employers and insurance companies, or politically biased.

Many other groups put forward opinions, including physiotherapists, ergonomists, management consultants, furniture manufacturers, chiropractors, and keyboard designers,

http://doi.org/10.18080/ajtde.v5n2.107

each promoting their own specialty as the main factor in diagnosing and solving the problem.

Some Observations about the Status of RSI

The following are notes resulting from a short and incomplete layman's survey of the history and status of RSI in the 30 years after the Telecom epidemic:

- the DAS/C incident in Telecom from 1983 has been frequently reported in articles and research about RSI, and appears a "landmark", unique around the world in magnitude and severity.
- Although keyboard related work has rocketed since 1983, no comparable incident appears to have been reported.
- Hocking's (1989) published research is frequently cited in articles and research papers about occupational health.
- The label for the complaint appears to be fluid. Prior to the Telecom incident, in the early 1980's one label used was “tendinitis", implying that the problem is "mainly inflammatory", a view that is often disputed.
- One emerging umbrella term appears to be “overuse injury (or syndrome)” or "occupational overuse syndrome" (OOS). Some different terms are used between countries; the Netherlands uses “complaints of the arms, neck and shoulders (CANS)” and the USA commonly uses “cumulative trauma disorder”, a term that includes back problems as well as RSI.
- a common set of symptoms are usually used to describe the condition. For example, "the condition mostly affects parts of the upper body, such as the neck, shoulder forearms, elbows, wrists and hands".
- The symptoms are said to range from mild to severe and usually develop gradually. They often include pain, aching or tenderness, stiffness, throbbing, tingling or numbness, weakness and cramp
- a common set of preventative actions and treatment are usually offered. These include changing the sitting position, mouse and keyboards settings, and posture. Treatments include rest, exercise, medications (anti-inflammatory, pain alleviators and some forms of antidepressants), physiotherapy and injections.
- There appears widespread agreement that in the vast majority of cases treatments are effective.

http://doi.org/10.18080/ajtde.v5n2.107

- For such a serious and recurring condition claimed to cause widespread work-related losses and suffering, there are minimal statistics on RSI complaints.
- There appears to be no standard definition for the syndrome in "official" occupational health and safety statistics. Statistics compiled in Australia and around the world usually do not list RSI as a separate condition. Data for neck, shoulder and hand, fingers and thumb injuries and diseases are often used as an indication of the problem and for information on workers' compensation claims for the condition.
- While there is agreement about the areas of the body affected, the symptoms and the treatments, there are opposing views about the link, if any, between repetitive hand use and physical injury - see the previous discussion.
- Despite the debate about the cause of the syndrome, there has been a steady increase in the cost of workers' compensation injuries.
- After some 30 years there still appears to be little scientific evidence that the DAS/C system and similar systems were the cause of the range of RSI symptoms listed.

So what caused the Epidemic?

Some of the main factors suggested as a cause of RSI are the workplace, the work station, the working conditions, the workload and the lack of worker satisfaction. Each of these is examined below.

In 1988, Booz Allen & Hamilton (BAH) (1988) conducted a benchmarking study comparing the performance of Telecom's DAS/C directory assistance service with similar services operating in four of the Regional Bell Operating Companies (RBOCs) in the USA. The results are shown in Table 4. Briefly, Telecom delivered a clearly inferior service at a significantly higher cost with RBOCs reporting minimal incidences of RSI. Differences are mentioned in the analysis.

The Workplace Centres were not a cause of RSI

The DAS/C centres provided far superior working conditions.

The quality of the old paper based workplaces varied around Australia, largely depending on the age of the centre. The metropolitan centres were well lighted and air conditioned, but were cluttered by the paper records which were heavy and awkward to use. All of the interiors of the DAS/C workplaces were new, professionally designed with lighting compatible with screen based work, open plan, and clean with relaxing colour schemes.

http://doi.org/10.18080/ajtde.v5n2.107

The Work Stations were not a cause of RSI

Telecom's IBM work stations were ergonomically designed with a “light touch” keyboard.

By 1982 there were over 2,000 workstations operating in computer-based directory assistance centres around the world supplied by IBM and other vendors similar to Telecom's DAS/C stations. By 1985 it is estimated that there were over 20,000 stations operating, mainly in the USA and Canada. In that year complaints about Telecom's DAS/C system with about 600 work stations reached a peak. Over the three-year period 1982 to 1985 there were minimal complaints of an RSI nature from the overseas centres.

The Frequency of Key Strokes on the IBM DAS/C Work Station was not the cause of RSI.

The DAS/C system required directory assistance operators to perform at a relatively low keying rate; an average of perhaps 20 keystrokes a minute with regular rest breaks.

A large number of occupations around the world performed keyboard-type work at consistently and far higher keying rates than the DAS/C system, many with shorter formal rest breaks and many with equipment that was not ergonomically designed. Examples include data entry keypunch operators producing Hollerith cards from the late 1800's and IBM cards from the late 1920's (in data entry centres); typists from the late 1800's (in typing pools); word processors from the 1960's and desk tops from the 1980's. The keying rate performed by Telecom's telegraphists was far higher with no RSI complaints until the arrival of the DAS/C system.

All of these various keyboard operators raised minimal complaints of an RSI nature over more than 100 years.

The Work Load was not a cause of RSI

The BAH (1988) benchmark study - see Table 4 - found that Telecom's directory assistance operators had a far lower work load than the RBOC operators.

During the year Telecom's operators worked only about 70% of the on-line hours per year of those worked by the RBOC operators, and during that much shorter time handled less than 70% of the calls per on-line hour. The RBOCs experienced minimal complaints of an RSI nature.

http://doi.org/10.18080/ajtde.v5n2.107

The Conditions of Employment were not the cause of RSI

The BAH benchmark study - see Table 4 - found that Telecom's directory assistance operators had far more generous conditions of employment than the RBOC operators, and on the evidence, a far less stressful existence.

Table 4: Some Results from the Booz Allen Hamilton Benchmarking Study - Directory Assistance - 1988

		Telecom	US RBOC's
Service Quality	**Calls Dropped**	**17%**	**less than 1%**
	Average Speed of Answer	**12 seconds(est)**	**3-6 seconds**
	Customer Perceived Quality - very satisfied	**86%**	**97%**
Working Arrangements	**Work Hours**	**34**	**37.5**
	Relief Time/Planned Breaks	**5.0**	**2.5**
	Unscheduled Breaks (hours)	**1.25 (est)**	**0.8**
	Effective on-line Time (hours/week)	**27.75**	**34.2**
	On Duty Hours (duty hours/total hours)	**82%**	**91%**
	Initial Roster Drawn (days in advance)	**90 days**	**14 days**
	Final Roster Adjustment	**7 days notice**	**½ hour in advance**
	Roster Traffic Basis	**one week average on past 4 weeks**	**weekly or daily traffic data**
	Use of Part Time/Casuals to manage peak loads	**10-15%**	**30%**
Working Days per Year	**Working Days in a Year**	**260**	**260**
	Less Vacation Days	**20**	**15**
	Sick Days	**15**	**5**
	Holidays	**13**	**10**
	Averaging Long Service Leave	**9**	**-**
	Averaging Days Off due to working on a Sunday	**4**	**-**
	Other Leave without Pay	**2**	**-**
	Average Working Days per Year	**197**	**230**

Output	Telecom	NY Tel	Pacific Bell	Bell Atlantic
Operators	**2,248**	**2,195**	**5,200**	**7,180**
Total Call Volume millions	**166**	**350**	**873**	**1,217**
On-line Hours /Operator/Year	**1,093**	**1,573**	**1,573**	**1,573**
Calls/On-line/Hr /Operator	**68**	**101**	**107**	**108**

Telecom's operators had a shorter working week (34 hours vs. 37.5), up to 33 less days worked per year (197 vs. 230), more relief time and planned breaks per day (5 vs. 2.5), more unscheduled breaks per day (1.25 vs. 0.8), and fewer effective on-line hours per week (27.75

http://doi.org/10.18080/ajtde.v5n2.107

vs. 34.2), more vacation days (20 vs. 15), more sick days (15 vs. 5), more public holidays (13 vs. 10), more advanced notice for rosters (90 days vs. 14), and enjoyed long service leave while the RBOC operators received none.

Also, Telecom operators, on average, were paid more than RBOC operators; while the base pay for Telecom operators was lower than the US, higher penalty rates caused a higher average cost per hour; $25.62 per hour for Telecom operators compared to $24.50 for the US. The union encouraged Telecom operators **to work non-busy hour shifts to earn penalty payments.**

Again, the RBOC operators raised minimal complaints of an RSI nature.

Lack of Work Satisfaction was not the cause of RSI

The DAS/C system provided a far higher work satisfaction than the old paper process.

As the quality of the paper-based service declined, callers became more frustrated and irritated when calls could not be connected, waiting times in the queue increased, searching times were too long, and when a number could not be provided. Searching the paper listings was physically arduous and often frustrating when the number requested could not be found.

The DAS/C system allowed the operator to provide a far higher quality of service; almost all calls were answered, call waiting times were markedly reduced, the time to find the number was far shorter, it was rare that the requested number could not be provided, the ergonomic work station required far less effort, fewer customers were frustrated and most callers were satisfied.

So what was the cause of RSI? A Sceptical Layman's View

Almost every worker feels a pain, aching or tenderness, stiffness, throbbing, tingling or numbness, weakness or cramp at some time during their working life.

Consider a Telecom directory assistance operator towards the end of 1982, the first year of deploying the DAS/C system. She (the majority of operators were women) has completed training and has been working on the DAS/C system. She has noted the number of positions in this DAS/C centre is less than the number of manual positions before, perhaps by up to 30%, and feels threatened.

She rises on a working morning and experiences a pain in her neck. She reflects that the pain may be caused by her new job on the DAS/C system; at this point there is no scientific causal evidence.

http://doi.org/10.18080/ajtde.v5n2.107

She visits her physician who asks what has she been doing lately? Has anything changed in her life that might have triggered this pain? Well, she has just started as a DAS/C operator and this involves repetitive keystrokes, an activity not needed in her previous work. The physician may consider this as a possible reason, and in any event, advises to take a few days off and a course of Aspros. At this point, there is no scientific causal evidence.

At work during next week the operator shares her experience of pain and the physician's diagnosis and advice. Some of her colleagues volunteer a similar experience. A union official, concerned about the steady fall in membership and the related effect on dues and income, has been assessing ways of obstructing deployment of the DAS/C system and the related fall in the operator workforce. The gossip among the operators is a blessing.

The union calls in the media and accuses Telecom of thoughtlessly and recklessly implementing a new system which injures the operators. At this point, there is still no scientific causal evidence.

The union's logic is that “because I have a pain, my new work must be the cause of my pain”, and this is reported by the media unchallenged. Other operators report pain and relate it to the new work. Yet more operators, whether they experience pain or not, realise the opportunity for workers compensation. The legal profession come to the same conclusion and adopts the allegation of cause. Other professions with an interest in occupational health and safety are also attracted including physicians, physiotherapists, and academics. The issue soon reaches the Labor government which intervenes on the side of the unions and compensation. The allegation of "injury caused by the work" becomes a tsunami beyond the rational.

The BAH (1988) benchmarking study was not available during the period 1982 -85. The less demanding life of the Telecom directory assistance operators compared to similar keyboard occupations in Australia and directory assistance operators in the USA was well known within Telecom but not documented. Telecom's management did not seek this comparative data and consequently could not present a sufficiently strong objective rebuttal. Ignored was the fact that some Telecom workers claimed compensation for suffering RSI symptoms under working conditions that had not changed during the same period; for example, word processors, telegraphists and clerical workers. Also ignored was the fact that some claimants did work with little repetitive element.

The alleged pain might have been a genuine condition for some operators but the lack of scientific data supporting the causal relationship had no influence on the mounting hysteria; the DAS/C system was the cause of an epidemic.

http://doi.org/10.18080/ajtde.v5n2.107

Conclusion

Most people experience pain at some time. It is clear that some have experienced pain caused by repetitive activity. The issue is how many and why?

From 1982 within Telecom the dominant concerns in the operator workforce was the operators' fear of job loss and the ATPOA's fear of a substantial fall in membership.

Pain reported by some directory assistance operators was attributed to the new DAS/C system by operators, the union, and contestable medical diagnoses by physicians and other professionals, with no objective evidence of a causal link of the injury to the work.

The claim that DAS/C caused the injury was sensationalised by the media and adopted by other employees, some members of the medical and legal professions, and some academics and others in the occupational health industry. These constituencies were undaunted by the fact that syndrome was occurring minimally, if at all, in similar workplaces elsewhere, and that occupations with far higher levels of keyboard repetition had been in existence for more than 100 years with minimal reports of injury.

The argument is that the union grasped the opportunity to ride the RSI phenomena as an industrial tactic to slow the deployment and limit the capability of the DAS/C system, negotiate an additional payout to redundancy, generate “injury” compensation, and at the same time slow the loss of members.

The Telecom epidemic was unique. It was likely triggered by a combination of factors; Telecom's monopoly which allowed the union to take action with no consequent loss of jobs (to competitors); the lack of strong leadership within Telecom's management to contest the unproven pain/cause allegation; the declining membership of the union; a Labor Government which could be relied on to support the union, and the opportunism and self interest of the media and some members of the professions.

In the 30 years since the Telecom epidemic there still appears little scientific basis for linking pain reported by directory assistance operators' with working on the DAS/C system.

References

Australian Telecommunications Commission. 1981-86. "Directory Services Review of Activities" for the years 1981/82, 1982/83, 1983/84, 1984/85 & 1985/86.

Australian Telecommunications Commission. 1975-86. "Annual reports for the years 1975/76, 1976/77, 1977/78, 1978/79, 1979/80 & 1980/81".

http://doi.org/10.18080/ajtde.v5n2.107

Booz, Allen, Hamilton. 1988. "A Benchmarking Study comparing the Directory Assistance Operations of Telecom and four Regional Bell Operating Companies."

Ewan, Christine; Lowy, Eva; and Reid, Janice. 1991. "Falling out of Culture': the effects of repetition strain injury on sufferers' roles and identity".

onlinelibrary.wiley.com/doi/10.1111/1467-9566.ep11340787/pdf.

Helliwell, P S; Taylor, W J. 2004. "Repetitive Strain Injury". Postgraduate Medical Journal, British Medical Journal. *pmj.bmj.com › Archive › Volume 80, Issue 946*

Hocking, Dr B. 1989. "Epidemiological Aspects of Repetitive Strain Injury in Telecom Australia". Medical Journal of Australia.

Patkin, Michael. 1989. "Problems of Computer Workers - Lessons from the Australian Debate". *www.mpatkin.org/ergonomics/rsi/montreal88.htm*

Patkin, Michael. 1991. "Limits to Ergonomics". A paper to the 1991 Annual Conference of the Ergonomics Society of Australia. *https://mpatkin.org/ergonomics/limits_erg_91.htm*

Quintner, J. L. 1994. "the Australian RSI Debate: stereotyping and medicine". *http://pudendalnerve.com.au/website/wpcontent/uploads/2013/09/The%20Australian%20RSI%20debate-%20stereotyping%20and%20medicine.pdf*

Rickertt, Jeffery. 2000. "Resistance on the Line - a history of Australian Telephonists and their Trade Unions, 1880-1988". PhD Thesis. School of History, Philosophy & Religion & Classics, University of Queensland. http://trove.nla.gov.au/work/34294972.

RSI and Overuse Injury Association of the ACT, Inc. 2017. *rsi.org.au.*

Thornthwaite, Louise. January, 1994. "Union Growth, Recruitment Strategy and Women Workers". Griffith University Series: Institute for Research on Labour and Employment. University of California , Berkley, USA.

http://doi.org/10.18080/ajtde.v5n2.107

U.S. Telco Industry History as a Prologue to its Future

Carol C. McDonough
University of Massachusetts Lowell

Abstract: The United States telco industry has been shaped by the interplay of technological advance, free enterprise, politics, public pressure, and government regulation. The history of the industry reveals a continuing tension between the forces of competition and concentration. Having coursed through eras of monopoly, competition, and regulated monopoly, the telcos are now in a more competitive arena. There is regulatory uncertainty on the issue of net neutrality.

Keywords: Net neutrality, technology, competition, monopoly, regulation

Introduction

The U.S. telco industry has been shaped by the interplay of technology, free enterprise, politics, public pressure, and government regulation. The history of the industry reveals a continuing tension between the forces of competition and concentration. High fixed costs and low and often declining marginal and average costs drive the industry toward higher levels of concentration. However, because a competitive market tends to benefit consumers and better allocate resources, the government has at times tried to promote a more competitive industry structure. Existing technology affects the relative validity of these two opposing forces. However, whether we examine the industry in its early stages, or today as it faces the issue of net neutrality, the basic tension between competition and concentration remains a pivotal challenge.

Having coursed through eras of monopoly, competition, and regulated monopoly, the telcos in the United States are now in a more competitive arena. As the telcos broaden their menu of services to include not only voice telephony but also television, Internet and wireless communication services, the telcos have competition from cable companies and wireless services. Going forward, the telcos face the challenge of successfully developing content and other information services rather than operating only as dumb pipes. The telcos' investment in landline infrastructure continues to be an important asset, because fixed wire provides backbone infrastructure for the rapidly-growing wireless industry.

http://doi.org/10.18080/ajtde.v5n2.108

Parts I and II discuss the history of U.S. telephony, with a focus on government regulation and market structure. Unlike in many European countries, the U.S. telco industry has been privately, rather than government, owned. Part I discusses the early years and the regulated monopoly era. Part II discusses the Post-Divestiture era, after the AT&T monopoly was separated into regional monopolies, and long distance and equipment segments. The conclusion contains a discussion about the future of the industry.

Part I

The era of patent-protected monopoly. The U.S. telco industry began as an unregulated duopoly in 1877. Both the American Bell Telephone Company and Western Union Telegraph built out competing and unconnected telephone service until 1879, when Western Union sold its network to Bell, and Bell agreed to exit the telegraph business.

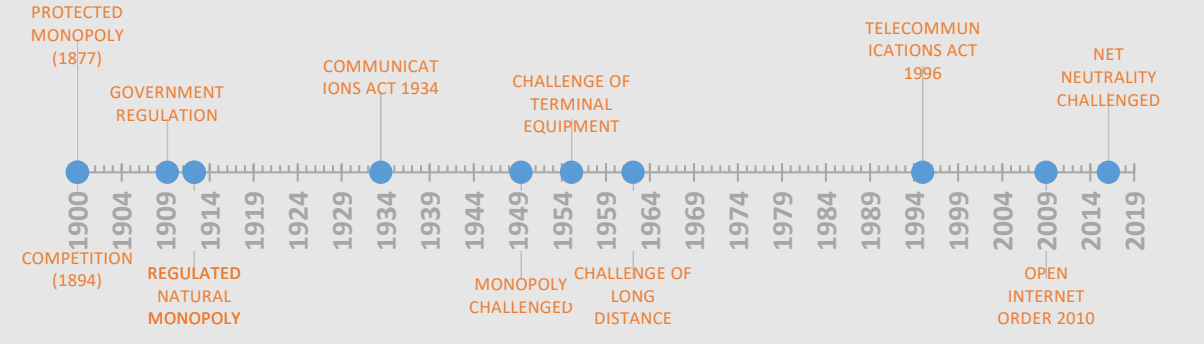


Figure 1 Telecommunications history in the United States

The Bell system then became an unregulated monopoly protected by patent rights. Bell licensed operating companies as territorial monopolies in major cities and metropolitan areas, requiring local exchange licensees to connect only with the Bell system (Horwitz, 1989). The Bell system initially licensed small independent companies to manufacture equipment using Bell patents. Before the patents expired and the equipment market opened to potential competitors, Bell vertically integrated by acquiring an interest in Western Electric Supply, the manufacturing branch of Western Union, in 1881. This acquisition acted as a barrier to entry and reduced potential competition. Western Electric was given a permanent exclusive license to manufacture telephones under the Bell patents. (Brock, 1981). American Bell created AT&T in 1885 to provide long distance service.

The era of competition. With the expiration of Bell's basic patents in 1893 to 1894, the industry became competitive, with widespread entry. By 1902, 3009 non-Bell commercial telephone systems had been established and by 1907, 49 percent of telephones were controlled by independent phone companies. (Brock, 1981).

http://doi.org/10.18080/ajtde.v5n2.108

To protect its market power, Bell attempted to extend its patent monopoly through broad patent interpretation or by combining multiple patents, winning more than 600 patent-infringement lawsuits (Horwitz, 1989). The Bell system also tried to retain market power by lowering prices and by connecting the operating companies with long-distance lines.

In 1899, AT&T acquired the assets of its parent, the American Bell Telephone Company. The Bell system had four major divisions: Long Lines, to interconnect local exchanges and long-distance services; equipment manufacture (Western Electric Company); research and development (Bell Labs); and Bell operating companies, to provide local exchange services.

As the industry grew, AT&T increased its market power by providing greater interconnectivity, a long-distance network, and in some cases, deeply discounted rates for service and equipment (Bornholz and Evans, 1983). Attempts by independents to form a long-distance network did not succeed. Bell's patents for long-distance transmission created a barrier to entry. As a result, the number of independent telephone companies began to decrease.

In 1907, AT&T began an aggressive campaign to buy out independent phone companies and to seek interconnections with strategically-located independent exchanges. AT&T asserted that telephony was a natural monopoly, with the slogan "One policy, one system and universal service." (Vail, 1907). AT&T argued that competing systems wasted resources; that, as subscriptions increased, network externalities increased the value of telephone service; and that a single company best provided integrated end-to-end service. The independents argued that competition reduced price and increased subscribers, therefore servicing consumers better than a monopoly system (Bornholz and Evans, 1983).

Government regulation begins. AT&T's attempt to unify the phone system led to lawsuits about rates and antitrust litigation. There was also movement toward government ownership, as existed in many European nations. Government regulation of the industry began with the Mann Elkins Act (1910), which gave the Interstate Commerce Commission (ICC) jurisdiction over interstate rates charged by phone companies. AT&T responded to the threat of government ownership by scaling back its initial plan of a unified system.

Telephony as a regulated natural monopoly. The industry's era as a regulated monopoly began with the Kingsbury Commitment of 1913. AT&T agreed to divest itself of its controlling stock holdings in Western Union, agreed to stop acquiring competing independent exchanges, and offered to open up its long-distance lines to independent exchanges that were more than fifty miles (80 kms) apart. In exchange, the U.S. Justice Department relaxed its antitrust pressure on AT&T. By reducing competition between Bell and the independents, the Kingsbury Commitment established a presumption of telephony as a local monopoly (Horwitz, 1989).

http://doi.org/10.18080/ajtde.v5n2.108

For a year (1918) during World War I, Congress placed AT&T under the control of the Post Office. This control, which ended because of concerns over rate increases, demonstrated the problems with government ownership. Instead, Congress passed the Willis-Graham Act (1921), which expanded federal authority by giving the ICC power to approve or disapprove consolidations and mergers of telephone and telegraph companies. However, telcos were exempted from antitrust regulation if the ICC determined that a proposed consolidation or merger would be "of advantage to the persons to whom service is to be rendered and in the public interest." The Willis-Graham Act effectively established telephone companies as natural monopolies that should be regulated and free from competition. As the U.S. House Committee stated, “there is nothing to be gained by local competition in the telephone business.” (Loeb, 1978).

Facing financial difficulties, many independent telcos supported the Willis-Graham Act: they wanted to be bought out by AT&T (Horwitz, 1989). The following year, AT&T and many independent phone companies signed the Hall Memorandum (1922). AT&T agreed not to purchase, or consolidate with, independent phone companies unless demanded for the convenience of the public or special reasons. The Hall Memorandum essentially defined the boundaries of the Bell System and the structure of the telephone industry by voluntary agreement. The early competitive era of telecommunications came to a close (Sterling, 2006). The industry structure stabilized in the early 1920s and remained largely unchanged until the 1960s, when interexchange competition re-emerged.

The Communications Act of 1934 transferred regulation of interstate telephone services from the ICC to the Federal Communications Commission (FCC) and consolidated regulation of wired and wireless service under the FCC. The FCC’s charge was to secure a rapid efficient nationwide wire and radio communication service at reasonable rates. Title II of the Act required telephone companies as common carriers to make services available to the public at reasonable rates.

At the time, AT&T had approximately 80% of the telephone market and the only significant long-distance telephone network. The remaining 20% was owned by a large number of scattered small local telephone companies dependent on the AT&T network for long distance service and interconnectivity (Brock, 1981). The creation of the FCC had the effect of supporting AT&T‘s monopoly position because the FCC accepted the institutional structure of monopoly in the telco industry (Horwitz, 1989). AT&T was allowed monopoly control over long distance voice telecommunications and to operate local monopoly telephone operating companies. In exchange, the FCC enforced common carrier legal obligations, requiring interconnection of carriers and charging carriers with the duty to serve all who requested service.

http://doi.org/10.18080/ajtde.v5n2.108

As a regulated monopoly, AT&T and the smaller telcos were guaranteed pricing levels that provided a fair rate of return. No criteria were set up to determine if rates were reasonable. Instead the Commission begin a process known as continued surveillance in which informal negotiations were carried on between the Commission and the telephone company to determine the appropriate interstate rates (Brock, 1981).

Monopoly challenged. After World War II, technological advances led to challenges to AT&T's monopoly position. The regulated monopoly model supported by FCC regulations had offered the benefit of universalizing telephone service, using value of service pricing and cost averaging to finance service expansion. However slow innovation, limited service options, and relatively high prices for heavy users created economic and technical incentives for large users to get out from under monopoly and for new providers to service these large corporate customers. In addition, the Justice Department resumed its pre-war investigation into AT&T's violation of antitrust laws.

Competitive challenges to AT&T's monopoly power arose because of two issues: terminal equipment attachments to the AT&T network and microwave technology for long-distance communication. These two challenges led to a modification of FCC regulations. It is debatable whether this shift signaled a reorientation of the FCC's regulatory policy toward introducing new competition into the telecommunications market, (Loeb, 1978) or was a response to unmet demand that the FCC reasoned would not have a significant negative effect on AT&T (Horwitz, 1989).

The Challenge of the Terminal Equipment Market. Incentives existed to enter the terminal equipment market at the end user level because AT&T equipment was priced above competitive costs and often did not meet specialized needs. In the 1940s and 1950s, the FCC repeatedly supported prohibitions against any foreign attachments to the AT&T telephone network. The Hush-A-Phone case (1956) challenged AT&T's ability to maintain absolute control over terminal equipment attachments. The U.S. Court of Appeals reversed the FCC's Hush-a-Phone decision stating "the mere fact that the telephone companies can provide a rival device would seem a poor reason for disregarding Hush-A-Phone's value in assuring a quiet line". The FCC subsequently issued an order requiring telephone companies to rescind tariff regulations that prohibited use of external devices, unless the device caused injury to the public or the telephone system.

Subsequently, AT&T's tariff prohibited the Carterfone device, a device that connected the landline system with two way mobile radios to provide radio telephony to entities located too far away from local networks, such as oil drillers. The FCC found that the AT&T tariff was unreasonable and discriminatory: it prohibited the Carterfone attachment whether or not it

http://doi.org/10.18080/ajtde.v5n2.108

harmed the telephone system, while permitting customers to attach equipment similar to Carterfone if the equipment was provided by AT&T. Large users, such as the American Petroleum Institute and the National Retail Merchants Association, had backed the liberalization of entry into the terminal equipment market (Horwitz 1989) and the U.S. Justice Department had found that the tariff against foreign attachments violated antitrust laws.

AT&T responded by requiring that all foreign attachments first connect to protective coupling manufactured and sold only by AT&T, in order to protect the integrity of the system. Although the protective device charges decreased competitors' profits, potential profits were still sufficiently high to attract new companies (Brock, 1981).

The FCC's Part 68 rule (1975) strengthened the impact of the Carterfone decision by requiring that terminal equipment connect to the network through standard plugs and jacks. Any manufacturer that met the Part 68 standard was authorized to produce customer premises equipment. Further strengthening the Carterfone decision were the FCC's Computer Inquiries, decided in the 1970s and 80s. In these Inquiries, the FCC promoted competition in the telephony equipment market by requiring telephone companies to sell equipment through separate subsidiaries. The Carterfone decision, and the Part 68 rule, may have helped to enable the growth of the internet. Without Part 68, users of the public switched network would not have been able to connect their computers and modems to the network (Ismail, 2011)

The Challenge of Long Distance Communications. During World War II, the U.S. government had invested heavily in microwave technology research, which was generally not patent-protected. However, microwave transmission required radio frequency spectrum allocations from the FCC. The FCC's initial policy was to grant regular microwave licenses only to common carriers and to grant non-common carriers only temporary licenses for experimental use. Even with the restricted FCC policies, many companies found it profitable to build temporary microwave networks because they were unable to obtain service from AT&T. In 1959, the FCC opened the market by agreeing to lease microwave frequencies to any private user, ruling that the availability of common carrier facilities would not be a factor in granting private microwave applications (Brock, 1981). This decision did not enable free entry into the long-distance market but did allow the build-out of private systems for proprietary use. AT&T responded with the Telpak tariff, offering large discounts for groups of private lines, to remove the financial incentive to build out a private system where AT&T facilities were available. However, the Telpak rate structure showed that AT&T's single line long-distance rates far exceeded microwave costs, creating an entry incentive.

In 1963, Microwave Communications Inc. (MCI) filed a request with the FCC for authorization as a common carrier. The FCC approved MCI's application in 1969, and a large number of

http://doi.org/10.18080/ajtde.v5n2.108

similar specialized microwave service applications followed. The applicants countered AT&T's opposition by arguing that they would offer specialty services not generally available at rates equal to or less than traditional companies such as AT&T. In 1971, the FCC ruled in favor of increased competition, stating that "there is a public demand for the proposed services and competition in the specialized communications field is reasonably feasible."

In 1974, MCI, facing financial challenges by offering only specialized communication services, filed with the FCC to offer Executive Network (Execunet) service, which effectively duplicated AT&T's regular message toll service and was a forerunner to MCI's entry into regular long-distance service. In 1975, the FCC rejected Execunet, claiming that MCI had been authorized to provide only private line or specialized communication service. In 1977, the U.S. Court of Appeals reversed the FCC decision, ruling that once the FCC licensed a firm to provide any service, that firm could provide every service unless the FCC specifically denied it. The Court questioned the legitimacy of AT&T's monopoly. "... There may be very good reasons for according AT&T de jure freedom from competition in certain fields; however, one such reason is not simply that AT&T got there first."

In 1980, the FCC issued its Computer II Order, which distinguished between basic services and enhanced services. Basic services, such as telephone service, which provided pure transmission capability, were regulated under Title II of the Communications Act. Enhanced services, that is, any service offering more than basic transmission over the telecommunications network (eg., computer processing applications) were not regulated under Title II.

Meanwhile, the Department of Justice's antitrust case against AT&T was gaining momentum. AT&T claimed that it was immune from federal antitrust regulation because it was regulated by the FCC. For the Justice Department, any settlement had to include opening up AT&T's network to competition (Sterling, 2006).

The landmark case against AT&T's regulated monopoly status was MCI's antitrust suit against AT&T. The case was decided in favor of MCI in 1980, with the jury and then the district court awarding MCI $1.8 million in damages. When AT&T filed a motion to dismiss, Judge Greene (1981) ruled that AT&T had violated antitrust laws "in a number of ways over a lengthy period of time ... and ...used their local exchange monopolies to foreclose competition in the...equipment market...and...monopolized the inter-city service market...." This suit, coupled with the Department of Justice antitrust suit brought against AT&T, eventually led to the voluntary breakup of the Bell System and a dramatic change in the structure of the telecommunications industry.

http://doi.org/10.18080/ajtde.v5n2.108

Part II

Divestiture of AT&T. In 1982, the Justice Department and AT&T agreed to settle the antitrust lawsuit brought by the Department of Justice. Under the terms of the Divestiture Agreement, on January 1, 1984, AT&T divested itself of its 22 local operating companies, which represented approximately two-thirds of AT&T's assets. These local operating companies were divided into seven regional Bell operating companies (RBOCs), which were subject to traditional rate of return regulatory oversight and were not allowed into competitive markets of long-distance, equipment manufacture, or data processing. AT&T became a vertically integrated company comprised of its extensive long-distance operations (Long Lines), Western Electric and Bell Labs. In addition, AT&T was permitted to enter emerging areas of communications, including the growing computer industry, a market not regulated by government.

The antitrust suit and the subsequent Divestiture Agreement signaled the industry that the government was committed to competition and may have had a dampening effect on mergers and acquisitions. In markets such as long-distance service where competition was possible, competitors developed new services and prices generally decreased. Where regulation was continued, primarily the local telco market, prices remained at comparable levels or increased (Shaw, 2001).

Post-Divestiture. The goal of the Divestiture Agreement was to subdivide the telephony industry into two sectors: a regulated monopoly sector for local telephony and a more competitive long-distance sector. However technological advances were creating opportunities for a more integrated telephony structure. Rather than accepting the confines of the Divestiture Agreement, the RBOCs began to file requests for permission to enter other lines of business such as equipment leasing and the provision of software programs. For example, Bell Atlantic, one of the RBOCs, set out to be a full-service company in the related emerging telecommunications and computer sectors. Because larger customers could potentially set up their own information systems, Bell Atlantic targeted medium-sized customers, offering this customer base everything from information services equipment and data processing to computer maintenance (Verizon Communications Inc., 2006).

During the 1990's, there were several major developments in the telecommunications industry that changed the nature and the structure of the industry. First of all, the convergence of the computer, broadcasting and telephone industries challenged the arbitrary distinctions in communication services, and the local telephone monopolies. Cable companies had the capacity to offer telephone service and telephone companies had the technology to offer television service.

http://doi.org/10.18080/ajtde.v5n2.108

Moreover, responding to increased demand for wireless communication services, in 1994, the FCC began to hold spectrum auctions, a competitive bidding process for the allocation of wireless spectrum licenses across the United States. While the FCC had granted experimental wireless system licenses in the 1970s and the 1980s, the auction process that began in 1994 (and is still ongoing) was the first opportunity for widespread competitive entry into the wireless market. It is interesting to note that during these first auctions, AT&T Wireless acquired spectrum in 94 percent of the Major Trading Areas (MTA).

The RBOCs responded to these market changes by merging with each other, and by acquiring wireless spectrum. By 2006, four of the seven original RBOCs had merged with or been acquired by the new AT&T Inc., which had directly or indirectly acquired both AT&T Corp. and AT&T Wireless. Two of the RBOCs had been acquired by Bell Atlantic, rebranded as Verizon.

During the two decades that followed the initial spectrum auction, AT&T and Verizon, with approval from the FCC and the Department of Justice, expanded their wireless footprint through mergers and acquisitions, and by successfully bidding on spectrum.

The Telecommunications Act of 1996, which modified the 1934 Act, had the goals of promoting competition and reducing regulation in the telecommunications industry. The Act redefined and deregulated telephone service. Prohibitions on equipment manufacture by the RBOCs were repealed. The 1996 Act significantly reduced the applicability of Title II common- carrier regulations by defining firms as providing either telecommunication services or information services. Telecommunication services were defined as the transmission between or among points specified by the user, of information of the users choosing, without change in the form or content of the information sent and received. Information services were defined as having the capability for generating, acquiring, storing, transforming, processing, retrieving, utilizing or making available information via telecommunications, and included electronic publishing. A telecommunications carrier, that is, a firm providing telecommunication services, was defined to be a common carrier. An information service provider was not required to meet common-carrier regulations. Congress thus created a significant regulatory distinction between companies that act as mere conduits for telecommunications and ones that manipulate or enhance communication. Congress also created a third category of information services, the telecommunications management exception, treating as a telecommunications service any use of an information service for the management or operation of a telecommunications service.

The 1996 Act also contained provisions to enable the build out of wireless communications. Section 332(c)(7) of the Act preserved state and local authority over zoning and land use

http://doi.org/10.18080/ajtde.v5n2.108

decisions for personal wireless service facilities, but preempted local decisions premised directly or indirectly on the environmental effects of radio frequency (RF) emissions, if the provider was in compliance with the Commission's RF rules.

The Challenge of an Open Internet: Regulatory Policy. Applying this statutory framework to broadband, in 1998, the FCC classified the transmission component of DSL, that is, the phone lines, as a telecommunications service. Then, in 2002, the FCC's Declaratory Ruling reversed position by reducing the applicability of Title II common- carrier requirements to broadband Internet providers. The Ruling held that cable modem service offered the end user more than data transmission capability and therefore qualified as an information service. This Ruling brought cable, DSL, and other wireline internet providers outside of the common carrier requirement. The Ruling was challenged by potential competitors seeking to use existing wireline networks. In 2005, the U.S. Supreme Court upheld the FCC's 2002 Ruling in National Cable and Telecommunications Association v. Brand X Internet Services, agreeing that a cable internet provider is an "information service", not a "telecommunications service." Thus, competing ISPs like Brand X Internet were denied access to the wires to provide competing internet service. The Supreme Court's approval solidified FCC authority to create a regulatory framework for ISPs under existing telecommunications laws. Moreover, by bringing ISPs outside of common-carrier regulations, the FCC signaled a policy shift that reflected the competitive and deregulatory intent of Congress (Hedge, 2006).

The next open-internet controversy occurred when Comcast claimed the management right to slow cable customers' access to a file-sharing service, because such programs consume significant amounts of bandwidth, and the FCC ruled against Comcast. Comcast filed a lawsuit and in April 2010, the U.S. Court of Appeals ruled that the FCC could not prevent internet service providers from blocking or slowing specific sites and charging video sites to deliver their content faster to users. The Court found that the FCC had not demonstrated that its action- "barring Comcast from interfering with its customers' use of peer-to peer networking applications- is reasonably ancillary to the …effective performance of its …duties" and had failed to cite any statutory authority that would justify its order compelling a broadband provider to adhere to certain open internet practices.

FCC Open Internet Orders. In response, the FCC adopted its first Open Internet Order (2010). This Order contained three rules governing internet service providers: no blocking, no unreasonable discrimination, and transparency. Restrictions on blocking and discrimination were subject to an exception for reasonable network management. The Order exempted mobile service providers from the anti-discrimination rule. The FCC's position on internet openness reflected its belief in the "virtuous circle" where richer and more diverse content on

http://doi.org/10.18080/ajtde.v5n2.108

the edge jumpstarts demand, which brings about infrastructure investment, which brings about even richer and more diverse content.

Verizon challenged this Order, arguing that the Order exceeded the FCC's regulatory authority and violated the Act. In 2014 the U.S. Court of Appeals upheld the FCC's determination that the 1996 Act granted the FCC authority to regulate private internet service providers and to enact open internet rules. However, the Court vacated the no blocking and anti-discrimination provisions because the FCC had chosen to classify broadband service as an information service under the Act, which expressly prohibits the FCC from applying common carrier regulations to such services. Likewise, the Court found that the no blocking rule applied to mobile broadband conflicted with the FCC's earlier classification of mobile broadband service as a private mobile service rather than a commercial mobile service.

In response to the Verizon decision, the FCC in 2015 (3-2 vote) amplified and clarified its 2010 Open Internet Order. The 2015 FCC Open Internet Order ruled in favor of net neutrality by reclassifying broadband as a common carrier under Title II of the 1934 Act and Section 706 of the Telecommunications Act of 1996. The Order set forth "bright-line" rules banning blocking, throttling (i.e., impairing or degrading lawful Internet traffic on the basis of content application service or use of a non-harmful device) and paid prioritization by providers of both fixed and mobile broadband Internet access service. The FCC's rationale for applying openness rules to mobile broadband was the widespread deployment and usage of 4G LTE mobile networks.

The U.S. Telecommunications Association et.al. filed a lawsuit challenging the FCC.'s power to classify Internet providers as common carriers under Title II, claiming that these rules will undermine future investment by large and small broadband providers, to the detriment of consumers. The lawsuit also questioned whether the FCC had the authority to group wired and wireless services under the same rules. In June 2016, the Appeals Court (2-1) upheld the legal authority behind FCC's Open Internet Order. Regarding the regulation of wireless services under the rules for wired services, one of the judges opined: "So if I'm walking in my house with an iPad, …at one end of the hall I connect to my Wi-Fi, at the other end, my device switches over to my wireless subscription — did Congress really intend these two services to be regulated totally differently even if I can't tell the difference?"

So far, no challenge with the Supreme Court has been filed. However, the newly-appointed FCC Chairman Ajit Pai opposes net neutrality, writing one of the two dissenting opinions in the FCC's 2015 Open Internet Order and calling net neutrality a last-century bit of regulation developed to tame a 1930's monopoly.

http://doi.org/10.18080/ajtde.v5n2.108

Conclusion

The telecommunications industry globally has been propelled by technological advance. The history of the U.S. industry shows that technology has triggered changes in the regulatory environment and the structure of the industry, from regulated monopoly to a more competitive market.

Historically, all three levels of government, legislative, executive and judicial, have contributed to the current regulatory framework. The FCC has issued rulings and orders, some of which have been upheld or overturned by the Appeals Court and the Supreme Court. Congress has enacted legislation, such as the Communications Act of 1934 and the Telecommunications Act of 1996. During periods of major transformation in the industry, FCC regulation has tended to lag behind technological advance. The Courts and Congress, often with pressure from industry, have introduced significant changes in industry structure and regulation.

Going forward, the impact of new technology on industry structure is unclear. Given a permissive regulatory environment, new technology that creates profit opportunities may attract new entrants, expanding competition. However, if technological advances introduce opportunities for increased economies of scale, the result may be a less competitive market, again given regulatory assent.

Currently, fixed-wire telecommunication companies have built out high-capacity, low latency, reliable fiberoptic and coaxial networks, which service end users and are the backbone of the wireless industry. At the end user level, these telcos face the challenge of operating as dumb pipes and/or operating in the competitive arena of content providers. As dumb pipes, the telcos provide core infrastructure, since streaming content from edge providers requires internet access and wired internet access has superior performance metrics. However, open internet regulation may significantly reduce the profitability of operating as dumb pipes.

As content providers, telcos compete with edge providers who have a jumpstart on knowledge of their consumer base and consumers' digital footprint. Telcos in the streaming market as content providers also face edge competitors who enjoy brand-name recognition and consumer loyalty. However, telcos have the advantage of being able to bundle streaming services with internet and/or cable TV and/or landline telephony. In the absence of net neutrality regulations, telcos could make it difficult and costly for start-up edge providers to function in the content market.

The structure of the U.S. telecommunications industry will be affected by the outcome of the net neutrality debate. Although the U.S. Appeals Court has supported the FCC ruling for net neutrality, opponents may challenge this ruling with the Supreme Court.

http://doi.org/10.18080/ajtde.v5n2.108

The argument against net neutrality is that such regulation creates uncertainty and reduces investment in broadband. ISPs oppose net neutrality, claiming that they have made considerable investment in fiberoptic landlines and should be able to obtain a reasonable return on their investment. Without the expectation of a reasonable rate of return on future infrastructure investment, ISPs say that they will not expand and improve the broadband network. Data from US Telecom shows a $1 billion decline in total U.S. broadband spending in 2015 (the year the FCC ruled for net neutrality) compared to 2014. Yoo (2014) found that the U.S. policy (at the time) of not regulating broadband as a public utility subject to common carrier requirements was more effective than the European net neutrality approach at reducing the digital divide. Open internet or net neutrality may also raise privacy concerns because government regulators need a mechanism to monitor, verify and enforce open internet regulation.

In support of net neutrality, startups claim that without such regulation, broadband providers could take advantage of their control over consumers' access to the internet by charging companies extra fees to reach customers faster, or making deals with preferred competitors, or by offering ancillary services themselves. Such abuse of market power is particularly likely in the numerous geographic areas serviced by only one high speed service. Moreover, there is a great deal of vertical integration in the industry, with major U.S. telcos owning or colluding with media conglomerates. The potential for abuse of market power among U.S. ISPs supports the argument for net neutrality regulation.

References

Bornholz, R. & Evans, D. (1983). The Early History of Competition in the Telephone Industry, in Breaking Up Bell, Evans, D. (ed.) North-Holland.

Brock, G. (1981). The Telecommunications Industry, Harvard University Press. U.S.A.

Comcast Corporation Petitioner v. Federal Communications Commission and United States of America, Respondents (2010). Argued January 8, 2010. Decided April 6, 2010.

Federal Communications Commission (1968). In the Matter of Use of the Carterfone Device in Message Toll Telephone Service, 13 2d 420.

Federal Communications Commission (1975). Part 68 Rule.

Federal Communications Commission (1980). Computer Order II, 384, 4200 (1980), 47 CFR 64.702.

Federal Communications Commission (1998). Memorandum Opinion And Order, And Notice Of Proposed Rulemaking FCC 98-188.

http://doi.org/10.18080/ajtde.v5n2.108

Federal Communications Commission (2002). In re Inquiry Concerning High-Speed Access to the Internet Over Cable and Other Facilities: Internet Over Cable: Declaratory Ruling and Notice of Proposed Rulemaking, 17 F.C.C.R. 4798.

Federal Communications Commission (2010). Open Internet Order.

Federal Communications Commission (2015). In the Matter of Protecting and Promoting the Open Internet, Washington, D.C. Adopted February 26, 2015, Released March 12, 2015.

Greene, H. (1981). Opinion (on Denial of Motion to Dismiss), 524 F. Supp. 1336 (September 11, 1981).

Hedge, J. (2006). "The Decline of Title II Common-Carrier Regulations in the Wake of Brand X: Long-Run Success for Consumers, Competition, and the Broadband Internet Market" Commlaw Conspectus, Vol. 14.

Horwitz, R. (1989). The Irony of Regulatory Reform, Oxford University Press.

Ismail, S. (2011). "Transformative Choices: A review of 70 years of F.C.C. Decisions," Journal of Information Policy 1

Loeb, G. (1978). "The Communications Act Policy Toward Competition: A Failure to Communicate," Duke Law Journal, Vol. 1978 March No. 1.

National Cable & Telecommunications Association v. Brand X Internet Services (2005). 125 S.Ct. at 2695-99.

Shaw, J. (2001). Telecommunications Deregulation and the Information Economy, Artech House.

Sterling, C, Bernt, P. & Weiss, M. (2006): Shaping American Telecommunications, Lawrence Erlbaum Associates.

Supreme Court of the United States (2005). National Cable &Telecommunications Association et al. v. Brand X Internet Services Et al. No. 04-277. Argued March 29, 2005-Decided June 27, 2005.

U.S. Congress (1934). Communications Act of 1934, June 19, 1934.

U.S. Congress (1996). Telecommunications Act of 1996, January 3, 1996.

U.S. Court of Appeals, D.C. (1956). Hush-A-Phone v. Federal Communications Commission and the United States of America, 238 F.2nd 266.

U.S. Court of Appeals, D.C. (2010). Comcast Corporation, Petitioner v. Federal Communications Commission and United States of America.). No.08-1291. Argued January 8, 2010. Decided April 6, 2010.

http://doi.org/10.18080/ajtde.v5n2.108

U.S. Court of Appeals, D.C. (2014). Verizon Corp. Petitioner v. Federal Communications Commission, 740 F. 3d 623, 2014.

U.S. Court of Appeals, D.C. (2016). United States Telecom Association, et al., v. Federal Communications Commission, No. 15-1063, U.S. Court of Appeals, D.C. Decided June 14, 2016.

U.S. Telecom (2015). Retrieved from https://www.ustelecom.org/broadband-industry/broadband-industry-stats/investment.

Vail, T. (1907). AT&T Annual Report.

Verizon Communications Inc. (2006) International Directory of Company Histories, Thomson Gale.

Yoo, C. (2014). U.S. vs. European Broadband Deployment: What Do the Data Say?, Penn Law Center for Technology, Innovation and Competition, June 2014.

http://doi.org/10.18080/ajtde.v5n2.108

Household bandwidth and the 'need for speed'

Evaluating the impact of active queue management for home internet traffic

Dr Jenny Kennedy
Media and Communication, RMIT University

Professor Grenville Armitage
Internet For Things (I4T) Research Laboratory, Swinburne University of Technology

Professor Julian Thomas
Media and Communication, RMIT University

Abstract: In this paper, we aim to contribute to the policy debate on bandwidth needs by considering more closely what happens in household networks. We draw upon both social and technical studies modelling household applications and their uses to show how queue management protocols impact bandwidth needs. We stress the impact of internet traffic streams interfering with each other, and describe three different categories of internet traffic. We demonstrate how the use of active queue management can reduce bandwidth demands. In doing so we consider how, and to what degree, household internet connections are a constraint on internet use. We show that speed demand predictions are skewed by a perceived need to protect the Quality of Service experienced by latency-sensitive services when using current gateway technologies.

Keywords: home broadband; bandwidth; domestic settings; Active-queue management; Internet traffic;

Introduction

One of the more contentious topics in Australian broadband policy is the present and future need for higher bandwidth services by Australian households. Broadband technologies are developing rapidly, with the emergence of a new, more complex ecosystem of domestic connections placing greater strain on household bandwidth requirements. Within the household ecology, mobile devices and Wi-Fi are now an alternative and complementary

http://doi.org/10.18080/ajtde.v5n2.99

mode of access for large numbers of Australians. The volume of connected devices and applications have proliferated both within and beyond households. This includes the nascent Internet of Things (IoT). Home broadband traffic is increasingly a mix of both latency-sensitive applications (such as online games and video conferencing services) and latency-tolerant applications (such as content streaming and data sharing services).

Increased bandwidth is attributed to increases in social welfare, GDP and consumption (Centre for Energy-Efficient Telecommunications, 2015; Deloitte, 2016). Increased bandwidth is also equated generally with social and economic benefits:

"increased bandwidth is accompanied by increased participation in the digital economy, in online activities, and in the use of entertainment and communication services and technologies; and that increased digital literacy emerges through experience and use of HSB [High Speed Broadband]" (Wilken et al, 2011, p. 10).

In this paper, we aim to contribute to the policy debate on bandwidth needs by considering more closely what happens in household networks. We draw upon both social and technical studies modelling household applications and their uses to show how queue management protocols impact bandwidth needs. We demonstrate how the use of active queue management can reduce the need for bandwidth. In doing so we consider how, and to what degree, household internet connections are a constraint on internet use. We show that speed demand predictions are skewed by a perceived need to protect the Quality of Service (QoS) experienced by latency-sensitive services when using current gateway technologies.

The contemporary networked household

To understand the bandwidth needs of households, we first have to consider the contemporary networked household in more detail. We can point to four critical trends: the proliferation of household connections; a combination of intensive and extensive growth in internet use; the emergence of new connected services; and increasing diversity in Australian households. Each of these four trends are significant in their own right and we consider them briefly below.

The proliferation of household connections

Australian households have an average of nine internet-connected devices; and 9% of households have 20 or more internet-connected devices (Telsyte, 2015). These numbers are predicted to rise rapidly in future years, with the average household having up to 29 connected devices by 2020 (Telsyte, 2015). Corporate research firm International Data Corporation (Turner, 2016) forecast 30 billion devices to be connected by 2020 globally,

http://doi.org/10.18080/ajtde.v5n2.99

while Gartner Inc (Gartner, 2015) report a prediction of 20.8 billion connected devices by 2020. At present, globally, around 5.5 million new devices are connected daily. Much of this dramatic increase is attributed to a proliferation of connected objects within the home, including white goods, health monitoring devices, and the IoT.

The combination of intensive and extensive growth in internet use

The proliferation of connected devices, combined with mobility, lead to both increased use, and concurrent use, which places greater demand on bandwidth services. Concurrent usage behaviours vary based on a number of contributing factors. For example, users are more likely to multitask if they have access to multiple connected devices (Hassoun, 2014). Connected general-purpose devices, such as PCs or laptops, are also likely to facilitate multiple connected applications at once. The IoT will vastly increase the number of connected devices in households, further impacting the average peak number of applications concurrently connecting to the home network.

The emergence of new connected services

There is a growing array of internet services: in Australia, one notable example is the emergence of video streaming services such as Netflix, which includes higher resolution content ('4K', or 'ultra high definition'). These services, while not necessarily bandwidth intensive in their own right, are examples of an increasing number of applications requiring some level of consistent bandwidth and increasing the aggregate demand. Another example is the growing popularity of online education, which makes use of a wide range of media, but particularly relies on video and audio. Telehealth and home automation are other emerging areas of service that require consistent bandwidth and increase aggregate demand. These examples of new media use require us to rethink common assumptions, such as the idea that demand for 'video' is mainly about entertainment.

The increasing diversity of Australian households

Considerations of household internet use need to take into account the range of living and working arrangements. The average household in Australia has 2.6 people (ABS, 2015).

To help illustrate variations in household formations, Telsyte (2015) proposes four simplified (and heavily urban-centric) household profiles, each with distinctive patterns of use. Though obviously limited, these profiles are productive for demonstrating that households utilise internet services differently at home. Each of the household types represent different models of social activity around internet use, characterised by varying patterns of peak concurrent application use based on household members and consumption patterns.

http://doi.org/10.18080/ajtde.v5n2.99

Table 1 Household profiles adapted from Telsyte (2015) report

Household profile	Peak app stack 2015	Predicted peak app stack in 2020
Dual professional households with children ("The Hectic Household")	12	19
Single or dual-parent households with children ("Suburban dreamers")	7	13
Couples without children ("City Living")	11	15
Shared living ("Shared living")	8	12

Contemporary household rhythms and peak app stacks

Rhythms of bandwidth consumption vary based on many factors. Patterns of internet usage cannot be assumed to be uniform given the known impact of socio-economic contexts (Thomas et al, 2016), infrastructure and geographical factors (Kennedy et al, 2017; Wilken et al 2013). Education, employment, social and entertainment practices are increasingly mixed, with activities overlapping both temporally and spatially within the domestic space (Nansen et al, 2009, 2010, 2011; Gregg, 2011). Increasing demands are placed on home bandwidth services and these demands intensify at peak times. What constitutes peak times for usage may vary and involve different configurations depending on the household and types of applications in use.

Kenny and Broughton (2014) identify that the crucial dilemma in future bandwidth use within the home is to do with anticipated peaks of concurrent application use (referred to as an 'app stack'). Kenny and Broughton identify four different types of application: primary (which they categorise as forms of streaming multimedia content, including TV, YouTube, HD video calls, and streamed and interactive games); secondary (content down and uploads, i.e. cloud storage, torrents, software and OS downloads, and non-HD video calls); web surfing; and finally, low-bandwidth traffic, which incorporate all other forms of internet usage. These four categories are characterised by user activity rather than application bandwidth specificities, e.g., Kenny and Broughton state "primary applications are those apps that are primarily used 'one at a time' by a given individual" (2014, p. 16). Subsequently, the combinations of concurrent application use depict rhythmic patterns of household activities and presume peak bandwidth use occurs during peak app stack. These categories give a broad overview of the types of applications people use concurrently and show how demands are driven less by the number of devices, and more by the types of applications and their particular bandwidth needs. Still, their categories of application overlook the impact of constrained bandwidth and latency sensitivities that also impacts bandwidth needs.

http://doi.org/10.18080/ajtde.v5n2.99

In the sections that follow we describe our own categorisation of applications based on latency sensitivity, and demonstrate how calculations of bandwidth requirements are less effective when they ignore the latency sensitivity of key applications.

User perceptions of internet Quality of Service, and impact on activities

Bandwidth needs are driven both by application requirements and by user perceptions of internet QoS or performance. A key contributor to QoS is Round Trip Time (RTT), meaning the time it takes for a signal or packet to travel from source to specific destination and back again. Lower RTT usually leads to better QoS.

Table 2 shows some typical RTTs in milliseconds (ms) that might be experienced by home-based applications when accessing remote services over an otherwise idle home gateway (also referred to as RTT $_{base}$). The RTT $_{base}$ is smallest when remote servers are closer.

Table 2: Typical figures for RTT $_{base}$ from Australia's east coast (Armitage & Heyde, 2012)

RTT base	Distance
10ms	Domestic, intra-ISP servers
40ms	Domestic, inter-ISP or off-net servers
180ms	Australia to US West Coast servers
240ms	Australia to US East Coast servers
340ms	Australia to European servers

RTTs are inflated when there are queuing delays (e.g. due to temporary buffering at congested network routers and switches along the path) and mutual interference between different categories of internet traffic in concurrent use. Such latency issues frustrate users (Ceaparu et al., 2004). Users are especially likely to report QoS issues when streaming content or conducting video calls, describing frustration with buffering and call dropout. Thresholds and latency tolerances are somewhat vague and arbitrary. Users are reported to have a threshold of up to 500ms for web page retrieval before becoming frustrated when searching for information (Arapakis et al., 2014) and to prefer sub-100ms for highly interactive first-person shooter (FPS) games (Armitage et al, 2006).

Inflated RTTs impact home internet activities whereby users adapt their practices to accommodate idiosyncratic tolerances of QoS. Strategies involve allowing more time to complete activities, varying and combining tasks, or considering alternate contexts, i.e. completing tasks at a different time, on a different device, or even a different network (Wac et al., 2011).

Users perceive that increased bandwidth will reduce RTTs and improve QoS. This perception is both fuelled by and reflected in market attention towards increased bandwidth at a cost of

http://doi.org/10.18080/ajtde.v5n2.99

other technological solutions. Increased bandwidth does not directly translate to reduced RTTs in the domestic context. RTTs vary heavily depending on particular household internet traffic streams which mutually interfere with each other. Bandwidth requirements depend on how often mutual interference between internet traffic is likely to occur, and consumer's tolerance of RTT inflation triggered by such mutual interference.

Categories of internet traffic

Mutual internet traffic interferences are experienced differently depending on the type of traffic. We identify three different categories in the section below.

Latency-sensitive and interactive traffic

Interactive applications involve steady streams of packets between two points on the network at intervals dictated (and limited) by the applications themselves. Some of these applications are continuously interactive in nature, involving a human at one or both ends of the network path, generating and consuming data sent over the network in real-time. Examples include:

- Voice over IP (VoIP)
- Multi-party voice/video conferencing (such as Skype, Facetime, and remote education services)
- Online games (particularly `twitch' games like First Person Shooters, or other highly immersive environments)
- Real-time, remote medical monitoring services

Other latency sensitive applications are sporadically interactive, generating short bursts of traffic infrequently, and benefiting from low RTTs for responses. Examples include cloud-based IoT and home automation services typified by nascent products such as Google's Home, Apple's HomeKit, Amazon's Alexa, and third-party programmable automation platforms such as `If This Then That' (internet.ifttt.com). The timeliness of information transfer is important for all latency sensitive traffic.

There are also behind-the-scenes, latency-sensitive activities. For example, when a user clicks on a new link on a web page their browser does a domain name system (DNS) lookup to determine the target site's actual IP address before retrieving the remote site's content. DNS lookups traverse the home's connection to their ISP, where delays due to interference from other traffic make web browsing more tedious and less responsive. Similarly, services that rely on transmission control protocol (TCP) connections (like web browsing, sending emails, starting new file downloads or uploads, and so forth) can feel far less responsive when a home's RTT to the outside world is inflated by congestion in the home gateway.

http://doi.org/10.18080/ajtde.v5n2.99

Latency-tolerant (elastic) traffic

Latency-tolerant traffic is elastic, in that the application is flexible in terms of RTT and tolerates slowing down or speeding up as dictated by available bandwidth.

Examples include:

- Photo and video sharing applications sync'ing content to/from “the Cloud”
- Web browsers retrieving embedded digital objects to render pages
- Sending and receiving emails with large attachments
- Peer to peer file transfer applications
- Remote/offsite backup systems
- Application update systems (such as triggered by Microsoft Windows updates, or iOS Android App Store updates)
- Downloading podcasts, movies, TV show episodes or music tracks in their entirety for later, offline playback
- Instant messaging / notifications with multimedia attachments (Twitter, Apple's iMessage, and so forth)

Elastic traffic is measured on time to completion (TTC) – how long to upload or download a podcast, photo or app update, and so forth. TTC goes up as the size of objects being transferred goes up and/or the bandwidth goes down. Consequently, bandwidth requirements for elastic applications depend on consumer tolerance for TTC, which in turn depend on the social context of application use. For example, one consumer might accept a TTC of 15 minutes when downloading a 60-minute video, while another consumer considers a TTC over 6 minutes to be unacceptable.

Elastic applications often use TCP as their underlying transport. Unless limited by the application itself, elastic traffic will consume as much bandwidth as TCP can extract from the network at any given instant. This causes home gateway congestion that increases RTT for all internet traffic sharing that gateway. TTC can degrade as RTT increases; new TCP connections take longer to start-up, and active TCP connections take longer to recover from regularly occurring packet losses (from brief slow-downs to multi-second stalls as competing TCP connections step in and briefly consume all of a path's capacity).

Streaming content traffic

Examples of streaming content include:

- internet TV and movie services (such as Netflix)
- internet radio services

http://doi.org/10.18080/ajtde.v5n2.99

Streaming services are initially interactive (and latency sensitive) when consumers are using online menus to select content. Once the selected content has begun playing (streaming), the service can adapt to variations in network latency.

Average bandwidth requirements can be estimated from the audio/video encoding rates of the content being streamed. However, the short-term behaviour of streaming traffic can have distinctly aggressive characteristics akin to repeated bursts of short-lived elastic traffic.

Services built on technologies such as DASH (Dynamic Adaptive Streaming over HTTP) will usually begin a stream by pulling down tens of seconds of content as fast as possible, then settling into a regular pattern of short bursts of data traffic as the client retrieves 'chunks' of content piece-meal from the server over time. In addition, DASH-like services will usually adapt the content quality (and hence size in bytes) of newly requested chunks depending on the speeds achieved while retrieving previous chunks. Commonly deployed on top of conventional TCP, such traffic causes periodic short bursts of congestion on the home broadband link as each new content packet is retrieved.

Combining the bandwidth needs of each category

Today's home internet users are likely to be dissatisfied by an internet service offering (a) high latencies for latency-sensitive applications, (b) long TTC for important elastic applications, and/or (c) poor streaming quality. They require sufficient downstream and upstream bandwidth to ensure satisfactory service during periods of mutual interference where applications from all three categories are simultaneously active.

Streaming traffic is an attractive category on which to base rough bandwidth estimates, as the average requirements can be estimated from the consumer's desired video and audio quality. For example, it is reasonably simple to estimate the number of MBytes it takes to stream TV or DVD-quality movie content per minute using common encoding and compression schemes. However, rather than flow smoothly, most streaming traffic hits the home's access link with short bursts of elastic-like traffic, consuming as much spare bandwidth as is available at the time during each burst.

Elastic applications are more complicated. They only require enough bandwidth to achieve an acceptable TTC. But by using TCP for transport, elastic applications will typically consume whatever extra bandwidth happens to be available at the time, leading to even shorter TTC than the user may need. What constitutes an acceptable TTC (and hence minimum bandwidth requirement) depends greatly on the application itself, usage-patterns and user tolerance thresholds.

http://doi.org/10.18080/ajtde.v5n2.99

Significance of queue management in home gateways

In the home, there is typically RTT inflation during heavy traffic loads due to conventional home gateways using the first-in-first-out (FIFO) queuing protocol.

The internet requires certain amounts of buffering (queue storage) in routers and gateways to absorb transient bursts of traffic. During periods of low (or no) congestion, buffers are mostly empty and packets pass through with minimal additional delay. However, during periods of congestion late arriving packets can experience additional queuing delays. When bulk data transfers cause long-lived or cyclical queue build up, the conventional FIFO queue architecture means all traffic gets backed-up inside the queue, and everyone experiences worst case RTTs.

RTT inflation is higher through gateways with more buffer space, and lower as access bandwidth goes up. This is a key reason why many users perceive higher bandwidth service to be the critical factor for a positive internet experience. Unfortunately, many gateway vendors provide excessive amounts of buffering in an attempt to minimise packet losses by TCP connections. Often referred to as 'bufferbloat' (Gettys & Nichols, 2011), the effect is to exacerbate RTT inflation during congestion. Dropped packets cause the higher layer TCP connection to slow down and retransmit the lost packet, which increases TTC not network layer RTT. RTT is inflated by having more queuing delay before the loss occurs.

We argue that traffic queues management is the most crucial factor when determining bandwidth needs. Downstream traffic queuing is managed at the internet service provider (ISP) end, whereas upstream traffic queuing is managed at the home gateway. Active Queue Management (AQM) at home gateways can reduce the upstream bandwidth required to support satisfying user experiences.

Motivated by concerns over 'bufferbloat', recent internet Engineering Task Force (IETF) interest has focused on AQM schemes such as Proportional Integral controller Enhanced (PIE) (Pan et al, 2013), Controlled Delay (CoDel) (Nichols & Jacobson, 2016) and FlowQueue-CoDel (FQ-CoDel) (Hoeiland-Joergensen et al, 2016). These AQM schemes provide burst-tolerant congestion signalling at far lower levels of queuing delay than is typical of classical FIFO (or tail drop) queue discipline. A summary of work studying the performance of different AQMs for a variety of traditional Internet applications can be found in Hoeiland and Joergensen (2015).

In particular, FQ-CoDel assigns different traffic flows to sub-queues, uses CoDel to manage each sub-queue, assigns relatively even bandwidth share between sub-queues, and briefly gives priority to newly-populated sub-queues. As a result, an FQ-CoDel bottleneck achieves

http://doi.org/10.18080/ajtde.v5n2.99

latency reductions, capacity sharing, and priority for low-rate or transactional traffic (such as DNS, TCP connection establishments and VoIP).

We use a controlled experimental testbed to demonstrate the positive impact of switching from FIFO to FQ-CoDel buffer management (Armitage et al., 2017). Figure 1 illustrates the RTT of a VoIP flow competing with an upstream elastic TCP flow with 1, 2, 3, 4 or 5Mbps upstream link speed. With FIFO, the RTT inflation is severe – over 1 second at 1Mbps and still over 250ms at 5Mbps. Using FQ-CoDel over the same range sees the VoIP flow experience RTTs well under 100ms.

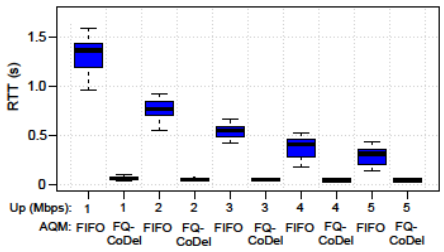


Figure 1: Significant RTT reduction when using FQ-CoDel instead of FIFO buffer management

Predictions of bandwidth need

To further demonstrate the significance of home gateway queue management for bandwidth requirements, we can emulate the probable internet usage scenario of a typical household. In doing so, we can give an estimation of the bandwidth demands for using FIFO queue management versus using AQM techniques.

The Hectic Household, as described by Telsyte (2015) averages 12 apps during peak time use. Drawing on ethnographic fieldwork on household rhythms of technology uses we can describe in some detail a realistic profile of such a household (Kennedy et al, 2015, 2017; Wilken et al, 2011, 2013).

Many applications start and stop throughout the day. We estimate that the peak app stack in this type of household of two parents and two children occurs in the early evening, at which time the following devices could be expected to be actively accessing the internet:

- Laptop #1
- Laptop #2
- Tablet #1
- Tablet #2
- Smart phone #1
- Smart phone #2
- FitBit

http://doi.org/10.18080/ajtde.v5n2.99

- Sonos speakers x3
- XBox
- Smart TV (or TV with set-top box)

Table 3: Application bandwidth requirements

Traffic Category	Application	Bandwidth requirements
Latency-sensitive applications	Online game	100 kbps down/up
	Video call	If audio-only call 100 kbps down/up Video & audio 200-500 kbps down/up
Latency-tolerant applications	Bulk downloads	Retrieving photos from `the cloud' / file downloads / receiving emails with attachments / software/firmware updates 5Mbps down
	Bulk uploads	Sync'ing photos or documents `to the cloud', sending emails with attachments 0.5Mbps up
	Short-lived TCP notifications	Social media / instant messages with attachments 0.2Mbps up/down
	Web browsing (with multiple windows open)	In Nov/Dec 2016 the average web page was 2.4MBytes of content spread over 106 HTTP requests across 35 different TCP connections. Requires both high speed (to download content) and low RTT (minimising TCP connection set up times) 3Mbps down
Streaming applications	HD video streaming	ABC iView @ high quality 1.5Mbps down Netflix SD@ 480p: 3Mbps down HD@720p: 5Mbps down Ultra HD (4K): 25Mbps down Stan SD: 3Mbps down HD (720p): 4.5Mbps down HD (1080p): 7.5Mbps down
	Audio streaming	Sonos 256 - 320 kbps down

In Table 3 we estimate the applications that could conceivably be in use simultaneously during the peak app stack, together with the typical bandwidth requirements of each application. Using these approximations, we can estimate the bandwidth required to provide uninterrupted and timely service for the household during this period. Adding the speeds required by each category in each direction separately gives us a downstream of 16.9Mbps (see Table 4) and upstream of 2.0Mbps (see Table 5).

http://doi.org/10.18080/ajtde.v5n2.99

Table 4: Downstream bandwidth requirements

Traffic Category	Total bandwidth requirements	Traffic
Latency-sensitive applications	0.6Mbps	0.1Mbps (game) 0.5Mbps (video call)
Latency-tolerant applications	8.8Mbps	5Mbps (bulk download) 4 x 0.2Mbps (concurrent notifications being received) 3Mbps (prompt web page retrieval)
Streaming applications	7.46Mbps	1.5Mbps (iView) 5Mbps (HD TV) 3 x 0.32Mbps (concurrent audio streams)

Table 5: Upstream bandwidth requirements

Traffic Category	Total bandwidth requirements	Traffic
Latency-sensitive applications	0.3Mbps	0.1Mbps (game) 0.2Mbps (video call)
Latency-tolerant applications	1.3Mbps	0.5Mbps (bulk upload) 4 x 0.2Mbps (concurrent notifications being sent)
Streaming applications	0.4Mbps	0.4Mbps (ACK traffic to support combined downstream streaming and latency tolerant download traffic)

A superficial conclusion would argue this household's needs could be met by a 18/2 Mbps service. However, this would fail to account for the RTT inflation of bulk transfer and streaming services hitting standard FIFO bottlenecks in the upstream and downstream directions. The household would notice significant degradation of latency-sensitive interactive activities during peak periods if the home were serviced at 18/2 Mbps with FIFO queue management.

In order to keep RTTs moderately bounded during bursts of elastic upstream traffic, this household requires at least 5Mbps in the upstream. Assuming the FIFO buffers aren't too long in the downstream direction, a hypothetical 18/5Mbps service might service this household with tolerable degradation from time to time. In the current market, they would need to purchase a plan of at least 25/5Mbps speeds.

Depending on personal thresholds, this household's needs could be catered for with 12/1Mbps speeds. But during peak app stack periods, while streaming services may remain arguably acceptable (albeit degraded), web browsing, bulk uploads and notifications would experience noticeably longer TTCs, and interactive applications would be regularly disrupted to the point of uselessness. The unsatisfactory solution for households today is to adapt their

http://doi.org/10.18080/ajtde.v5n2.99

usage patterns and eliminate simultaneous use of latency sensitive, latency-tolerant and streaming applications. However, this solution is increasingly ineffective as more unattended applications on tablets, phones and IoT devices launch bursts of elastic traffic at unpredictable times throughout the day.

Household requirements with AQM

When considering the household's overall bandwidth requirements if their broadband service used FQ-CoDel for active queue management in both directions, the speeds required by each category of traffic in each direction are the same as above. What changes is the overall bandwidth required from the ISP to minimise mutual interference.

FQ-CoDel isolates the different traffic flows from each other and keeps RTTs low. The household's latency sensitive traffic is protected and low RTTs are experienced, regardless of elastic and streaming traffic. In this scenario, a hypothetical 18/2 Mbps service becomes conceivable.

FQ-CoDel is also beneficial to elastic and streaming applications, as the underlying round-robin scheduling gives even sharing of available capacity during peak app stack periods. Compared to FIFO scenarios, with FQ-CoDel it is far less likely for one elastic traffic flow to impact capacity of other flows for periods of time. Consequently, the household would perceive the TTCs of various activities to be more consistent throughout the day.

Even 12/1Mbps speeds could be a realistic option if the household was willing to accept a modest degradation in TTCs for elastic applications and reduced streaming quality during peak app stack periods. During peak periods, the interactive traffic is protected (due to its low speed demands) and the remaining upstream and downstream capacity is divided equally between all other competing traffic.

The bandwidth requirements of most current latency-sensitive applications are dwarfed by the requirements of streaming media or elastic applications with demands for low TTC. However, it only takes one latency-sensitive application to be impacted by RTT inflation for the household to perceive their broadband service to be inadequate. Consequently, demand for significant (e.g. greater than 5Mbps) upstream bandwidth will exist whether we envisage one VoIP call during peak app stack periods, or multiple VoIP calls and multiple online games. Alternatively, we deploy an AQM (like FQ-CoDel or similar) at least in the upstream direction of home gateways to mitigate RTT inflation.

While the Telsyte report (2015) provides some alternative models for household populations (and hence peak app stack), there remains a need for greater insight into TTC tolerances and expectations of typical consumers for activities such as sending multi-media notifications,

http://doi.org/10.18080/ajtde.v5n2.99

sync'ing content to and from portable devices, and so forth. It is also important to constrain overly optimistic estimates by recognising limits imposed by the physical context of typical households. For example, the number of living areas, bedrooms, and so forth puts practical upper bounds on the number of streaming or internet access devices operating concurrently at any given time. The physical sizes of viewing areas also limits the size of practical TV screens (and hence, the degree to which a household may be satisfied with combinations of SD, HD and/or ultraHD streaming).

AQM deployment challenges

A number of practical issues mean that AQMs will take time to deploy. In the long term, equipment on either side of the broadband link between homes and their ISPs need to be upgraded. In the short term, significant benefits will accrue simply from having home gateways incorporate something like FQ-CoDel to manage buffers on the upstream side of their broadband service port(s).

ISPs typically use equipment from experienced data networking device vendors to control their end of the access link. However, upgrade schedules are guided by existing multi-year equipment or software contracts, and their vendors' willingness to incorporate AQM technologies in the medium to long term.

In contrast, home gateways may be owned and supplied by the ISP, or independently purchased by the consumer from a range of low-margin vendors. This raises several challenges.

ISP-owned gateways might be upgraded to support FQ-CoDel through automated, remote firmware update and reconfiguration. But this presumes the gateway's vendor plans to release new firmware with FQ-CoDel or similar AQM, and that the existing firmware allows remotely-controlled firmware updates. Well-known brands have been inconsistent in supplying updates to devices sold more than a few years earlier. This may well continue in regard to adding AQM.

Technically-savvy consumers might be motivated to try upgrading their own gateway's firmware. But only a small subset of consumers are likely to explore this option, constrained by reliance on new vendor-supported firmware or third-party open-source firmware that supports AQM.

A significant fraction of basic consumer gateways retail for less than $100 and are based around embedded versions of the Linux 2.x series kernel that do not support FQ-CoDel. If late 3.x or current 4.x series Linux kernels are on a vendor's product development roadmap, supporting FQ-CoDel in future models is easy. If a firmware upgrade of their existing FIFO-

http://doi.org/10.18080/ajtde.v5n2.99

based gateway is not an option, consumers face a choice between replacing their existing gateway with a new sub-$100 AQM-capable unit, or paying monthly for a higher downstream/upstream speed tier from their ISP. The former is likely to be a very attractive option for people whose downstream speed requirements are already met by one of the lower speed tiers.

Conclusions

Users and ISPs are in a position to leverage AQM technologies in the upstream direction of home gateways, independent of the political and market forces driving country-wide upgrades to broadband access services. This benefit is currently under-tapped and represents a niche market opportunity. Deployment of FQ-CoDel (or similar) would provide significant improvement to consumer experience of interactive services for those on a 12/1Mbps speed tier or ADSL2+ plans. Even with 25/5Mbps and higher tiers available, many households may find their needs for VoIP, games and casual internet use met by a 12/1Mbps plan coupled with FQ-CoDel.

Acknowledgements

This work was supported in part by a 2014-2016 Swinburne University of Technology / Cisco Australia Innovation Fund project titled “An evaluation of household broadband service requirements for educational innovation and Internet of Things”.

References

ABS. (2015). ‘3236.0 - Household and family projections, Australia, 2011 to 2036’. Australian Bureau of Statistics. Available from: http://internet.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3236.0Main%20Features42011%20to%202036

Arapakis, I., Bai, X., & Cambazoglu, B. B. (2014, July). ‘Impact of response latency on user behavior in web search’. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retriev*al (pp. 103-112). ACM.

Armitage, G., Claypool, M., & Branch, P. (2006). *Networking and online games: understanding and engineering multiplayer Internet games*. Cambridge: John Wiley & Sons. U.S.A.

Armitage, G. & Heyde, A. (2012). ‘REED: Optimising First Person Shooter Game Server Discovery using Network Coordinates’. *ACM Transactions on Multimedia Computing, Communications and Applications* (TOMCCAP), 8 (2). Available from: http://dx.doi.org/10.1145/2168996.2169000

Armitage, G., Kennedy, J., Nguyen, S., Thomas, J. & Ewing, S. (2017). 'Household internet and the 'need for speed': evaluating the impact of increasingly online lifestyles and the internet of things'. *Centre for Advanced Internet Architectures*, Technical Report 170113A. Available from: http://caia.swin.edu.au/reports/170113A/CAIA-TR-170113A.pdf

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., & Shneiderman, B. (2004). 'Determining causes and severity of end-user frustration'. *International journal of human-computer interaction*, 17(3), 333-356.

Centre for Energy-Efficient Telecommunications. (2015). 'Economic benefit of the National Broadband Network'. Centre for Energy-Efficient Telecommunications. Available from: http://www.ceet.unimelb.edu.au/publications/ceet-economic-impact-nbn.pdf.

Deloitte. (2016), 'Australia's Digital Pulse 2016'. Deloitte Access Economics, for Australian Computer Society. Available from: https://www.acs.org.au/content/dam/acs/acs-documents/PJ52569-Australias-Digital-Pulse-2016_LAYOUT_Final_Web.pdf

Gartner. (2015). 'Gartner says 6.4 billion connected "things" will be in use in 2016, up 30 percent from 2015'. Gartner Inc. 2015. Available from: http://www.gartner.com/newsroom/id/3165317

Gettys, J. & Nichols, K. (2011). 'Bufferbloat: Dark Buffers in the Internet'. *Queue*, 9 (11), pp. 40-54.

Gregg, M. 2011. *Work's intimacy*. Cambridge: Polity Press. U.S.A.

Hassoun, D. (2014). 'Tracing attentions: Toward an analysis of simultaneous media use'. *Television & New Media*, 15(4), 271-288.

Hoeiland-Joergensen, T., Hurtig, P. & Brunstrom, A. (2015). 'The Good, the Bad and the WiFi: Modern AQMs in a residential setting'. *Computer Networks*, 89, pp. 90-106.

Hoeiland-Joergensen, T., McKenney, P., Gettys, J., & Dumazet, E. (2016). 'The FlowQueue-CoDel Packet Scheduler and Active Queue Management Algorithm'. IETF Draft, March 18. Available from: https://tools.ietf.org/html/draft-ietf-aqm-fq-codel-06.

Kennedy, J., Nansen, B., Arnold, M., Wilken, R., & Gibbs, M. (2015). 'Digital housekeepers and domestic expertise in the networked home'. *Convergence*, 21(4), 408-422.

Kennedy, J., Wilken, R., Nansen, B., Arnold, M., & Harrop, M. (2017). 'Overcoming the tyranny of distance? High speed broadband and the significance of place'. In Griffiths, M. & Barbour, K. *Making Publics, Making Places*. Adelaide: University of Adelaide Press.

Kenny, R. & Broughton, T. (2014). 'Domestic bandwidth requirements in Australia: A forecast for the period 2013-2023'. Communications Chambers, for The CIE. Available from:

https://www.communications.gov.au/sites/g/files/net301/f/Forecasting-Australian-Per-Household-Bandwidth-Demand-Commun.pdf

Nansen, B., Arnold, M., Gibbs, M. R., & Davis, H. (2009). 'Domestic orchestration: Rhythms in the mediated home'. *Time & Society*, 18(2-3), 181-207.

Nansen, B., Arnold, M., Gibbs, M., & Davis, H. (2010). 'Time, space and technology in the working-home: an unsettled nexus'. *New Technology, Work and Employment*, 25(2), 136-153.

Nansen, B., Arnold, M., Gibbs, M., & Davis, H. (2011). 'Dwelling with media stuff: latencies and logics of materiality in four Australian homes'. *Environment and Planning D: Society and Space*, 29(4), 693-715.

Nichols, K., & Jacobson, V. (2016). 'Controlled delay active queue management'. IETF Draft, December 22. Available from: https://tools.ietf.org/html/draft-ietf-aqm-codel-06

Pan, R., Natarajan, P., Piglione, C., Prabhu, M. S., Subramanian, V., Baker, F., & VerSteeg, B. (2013). 'PIE: A lightweight control scheme to address the bufferbloat problem'. In *High Performance Switching and Routing (HPSR)*, 14th International Conference (pp. 148-155). IEEE.

Telsyte. (2015). 'Internet Uninterrupted: Australian Households of the Connected Future'. Telsyte, for NBN Co. Available from: http://www.nbnco.com.au/content/dam/nbnco2/documents/Internet%20Uninterrupted%20Australian%20Households%20of%20the%20Connected%20Future.pdf

Thomas, J., Barraket, J., Ewing, S., MacDonald, T., Mundell, M. & Tucker, J. (2016). 'Measuring Australia's Digital Divide: The Australian Digital Inclusion Index 2016'. Swinburne University of Technology, Melbourne, for Telstra. Available from: https://digitalinclusionindex.org.au/wp-content/uploads/2016/08/Australian-Digital-Inclusion-Index-2016.pdf

Turner, V. (2016). 'The internet of things: Getting ready to embrace its impact on the digital economy'. International Data Corporation. Available from: https://www.idc.com/getdoc.jsp?containerId=DR2016 GS4 VT

Wac, K., Ickin, S., Hong, J. H., Janowski, L., Fiedler, M., & Dey, A. K. (2011). 'Studying the experience of mobile applications used in different contexts of daily life'. In *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack* (pp. 7-12). ACM.

Wilken, R., Arnold, M., & Nansen, B. (2011). 'Broadband in the home pilot study: suburban Hobart'. *Telecommunications Journal of Australia*, 61(1), 5-1.

Wilken, R., Nansen, B., Arnold, M., Kennedy, J., & Gibbs, M. (2013). 'National, local and household media ecologies: The case of Australia's National Broadband Network'. *Communication, Politics & Culture*, 46(2).

http://doi.org/10.18080/ajtde.v5n2.99

Social Network Behaviour Inferred from O-D Pair Traffic

Mostfa Albdair

Faculty of Engineering, Misan University, Iraq

Ronald G. Addie

School of Agricultural, Computational and Environmental Science,

University of Southern Queensland, Australia

David Fatseas

School of Agricultural, Computational and Environmental Science,

University of Southern Queensland, Australia

Abstract: Because traffic is predominantly formed by communication between users or between users and servers which communicate with users, network traffic inherently exhibits social networking behaviour; the extent of interaction between entities – as identified by their IP addresses – can be extracted from the data and analysed in a multiplicity of ways. In this paper, Anonymized Internet Trace Datasets obtained from the Center for Applied Internet Data Analysis (CAIDA) have been used to identify and estimate characteristics of the underlying social network from the overall traffic. The analysis methods used here fall into two groups, the first being based on frequency analysis and second method being based on the use of traffic matrices, with the latter analysis method being further sub-divided into groups based on the traffic mean, variance and covariance. The frequency analysis of origin, destination and O-D Pair statistics exhibit heavy tailed behaviour. Because the large number of IP addresses contained in the CAIDA Datasets, only the most predominate IP Addresses are used when estimating all three sub-divided groups of traffic matrices. Principal Component Analysis and related methods are applied to identify key features of each type of traffic matrix. A new system called *Antraff* has been developed by the authors to

http://doi.org/10.18080/ajtde.v5n2.106

carry out all the analysis procedures.

Keywords: Social Network, Origin–Destination, Traffic Matrix, Principal Component Analysis.

Introduction

Because traffic data provide quantitative measurements of the intensity and quantity of communication between the entities (predominantly users and servers) which share access to a network, we should be able to estimate social network behaviour from traffic samples (trace files). The key difference between social network behaviour, as an interpretation of traffic, vs traditional traffic theory, is that it emphasizes end-to-end patterns, rather than traffic behaviour over time. For example, in traditional traffic theory, it is conventional to view a flow as equivalent, or a generalization, of a TCP flow. Thus, it has a start and an end. But in the analyses undertaken here, the start and end are not considered important. All the traffic between a certain originating IP address and a different destination IP address are viewed as a flow.

Social network behaviour is about who is talking to who, how much they are saying, and the patterns within these conversations. There are many potential patterns, so we need to focus on certain key patterns initially, for example, whether there exist groups of correlated sources, or of destinations, or of O-D flows. For reasons of privacy, the true identity of the users represented in a traffic trace is not normally available. In any case, the term *social network behaviour* is not used to refer to the identification of specific connections or relationships between users. In this paper *social network behaviour* is defined to mean *statistical characteristics* which result from the *variation in connection and quantity of communication between the members* of a community. An example of social networking behaviour which might be expected, and which is discovered in the traffic traces in this study, is the heavy-tailed character of the distribution of the total traffic between each origin-destination pair (O-D pair); that is to say: a small number of O-D pairs carry a large quantity of traffic, and a large number of O-D pairs carry relatively little traffic. This heavy-tailed behaviour is not just observed to occur, but its specific "shape" can, and has been, estimated.

We seek to interpret the traffic, even on one link, as a guide to the behaviour of the whole network. If the observed link is in the core network, the range of IP addresses which appear on it will be very large, and although the traffic observed is a sample, because the features being observed are of a statistical character, rather than concerned with specific user identities, there is no obvious reason for the use of a sample to cause the estimates obtained to be biassed.

http://doi.org/10.18080/ajtde.v5n2.106

Why social behaviour is important

Given the importance of modelling the traffic matrix (TM) for number of network aspects, good models of traffic matrices are important. The paper (Erramill, Crovella, & Taft, 2006) is concerned with the paucity of studies which are focused on complete models for TM, despite the importance of TM modelling. Traffic modelling and analysis is an essential preliminary step in network design and management. A number of point-to-point traffic models which have been studied in the literature are discussed in (Chandrasekaran, 2009), (Vardi, 1996). However, there is much less work currently available on models of all the traffic on a network.

In order to develop QoS architecture within the Internet networks, a strong traffic model is required for accurate performance evaluation and traffic analyses (Adas, 1997). What is needed is not only a model for the traffic on each link, but a model for the traffic of the whole network (Lakhina et al., 2004). This is the type of model we seek in this paper.

A whole network traffic model will enable to us to see a clear picture for the behaviour of traffic which is statistically sound. Whole network traffic analysis is difficult because the number of origins, destinations, and origin-destination pairs which arise may be so large that naive methods of analysis may fail, or the time taken to produce a result may take too long. In some cases there may be more than 1 million distinct sources, destinations, or O-D pairs referenced in a sample. On the other hand, in our analysis it may be necessary to identify correlations *between* different origins, or destinations, or O-D pairs, which will not be possible without some carefully designed algorithms.

To gain a better understanding of whole-of-network traffic it is useful to focus on origin-destination flows (O-D flows). We need to compare and contrast different origins, different destinations, and different O-D flows. This analysis needs to be statistical in its framework, and if possible we would like to identify the key statistical features which explain network behaviour in terms of origins, destinations, and O-D flows.

Source of data

In this paper, traffic data from Center for Applied Internet Data Analysis(CAIDA) (*Center for Applied Internet Data Analysis*, 2016) is analysed by a variety of methods to discover features of the social network from which the traffic is generated. The IP addresses in CAIDA datasets are anonymized. The Crypto-PAn tool was used to anonymize the IP addresses in the dataset. Crypto-PAn is a cryptography-based sanitization tool for network trace owners to anonymize the IP addresses in their traces in a prefix-preserving manner (Fan, Xu, Ammar, & Moon, 2004). This tool has the following properties: (a) the IP address anonymization is prefix-preserving. In other words, if two original IP addresses share a k-bit prefix, their anonymized mappings will also share a k-bit

http://doi.org/10.18080/ajtde.v5n2.106

prefix; (b) the same IP address in different traces are anonymized to the same address, even though the traces might be sanitized separately at different time and/or at different locations.

Table 1 lists all the CAIDA datasets which are used in this paper. This table also contains the figures in which results from that particular data sets.

Summary of paper

In Section , the traffic analysis methods are discussed. In Section , *the Antraff* software is described. In Section a series of experiments are described and discussed. Finally, the conclusion of overall work will presented in Section .

Traffic Analysis Methods

Frequently used IP addresses

A traditional *traffic matrix*, which was used for many years in telecommunication networks, contains, for each origin-destination pair, the "volume" of traffic (in Erlangs, or bits/sec), between the origin and the destination. By assigning a common column and row index to each node in the network, such data can be readily presented as a matrix. However, when monitoring Internet traffic, without complex processing to translate the IP addresses into node identities relative to a specific network, it may be difficult to identify the origin and destination *nodes* of packets. Using IP addresses in place of origins and destinations seems the logical step to overcome this problem, but the total number of IP addresses, even in IPv4, is too large to allow matrices with one row and column for each IP address.

For this reason, in order to analyse the traffic in a capture file such as those considered in this paper (*The CAIDA UCSD Anonymized Internet Traces 2014 - [20140320]*, n.d.), an essential first step is to identify the *frequently used IP addresses*.

Finding the relatively small number of IP addresses which occur a significant number of times in a capture file has been achieved as follows:

The total number of distinct IPv4 addresses is approximately 4 billion, hence even though IP addresses are integers, we can't use them as the indices of a vector or matrix. Even in quite a short traffic trace file, the total number of distinct IP addresses can easily exceed 1,000,000, and so it is not feasible to construct and manipulate vectors or matrices to describe the statistics of traffic from one IP address to another, even if this is precisely our objective.

http://doi.org/10.18080/ajtde.v5n2.106

So, instead, when matrices indexed by IP addresses are needed, instead of the IP address itself, as the index, we use its *rank*, and instead of including *all* IP addresses as potential indices, we only use the *most important*. IP addresses are all ranked, by their importance, measured as number of bytes sent from or to that address, and this rank is used both to select which IP addresses need to be included in the traffic matrix and the order in which they are included.

Traffic matrices

The simplest and traditional way to collect and manipulate traffic data is to create a matrix indexed by origins and destinations which contains the average bytes per second of the data from the origin to the destination, during the period of time considered. We will call this the *mean traffic matrix*. We can apply SVD to this matrix directly.

The mean traffic matrix can introduce some undesirable smoothing of the traffic. In particular, if the traffic is very bursty, the mean traffic matrix will not record this feature, and relying on it alone will lead to under-provisioning. We can avoid this by using a slightly different matrix. In this second approach, and the third, it is necessary to subdivide the traffic into time-slots during the analysis. The choice of time-slot length is significant, and will be varied. A typical value in the analyses presented here is 0.1 seconds.

When forming the second traffic matrix, which we term the *variance matrix*, the *square* of the bytes between each origin and destination, in each interval, is accumulated in each entry of the matrix, and divided by the interval length. We can also apply SVD to this matrix directly.

There are two variants of this type of matrix: if the mean-squared of the traffic in each interval is subtracted from the mean sum-of-squares, we arrive at the traditional variance, of the traffic passing between each origin and destination. However, it is also useful *not* to subtract the mean-squared. This matrix, which we could perhaps term the *power* matrix, includes the mean-squared, and hence is a better measure of the end-to-end load.

The third matrix type, which is the type used in (Lakhina et al., 2004) is formed by using O-D pairs, rather than origins and destinations, as the indices of the rows and columns. In each time-slot, the bytes delivered between each origin and destination are determined, and the *product* of entry C_{ij} in the covariance traffic matrix for this time-slot is the *product* of the bytes sent between O-D pair i and O-D pair j. These are then averaged over the trace to form the covariance traffic matrix.

Although these three traffic matrices may seem rather different, it turns out that they all reveal similar features of the traffic, and this will be explained in Subsection .

http://doi.org/10.18080/ajtde.v5n2.106

Principle Component Analysis and Singular Value Decomposition

Principal component analysis has been intensively studied and is widely in applied statistics. Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a new dataset. It's often used to make data easy to explore and visualize. The paper (Ringberg, Soule, Rexford, & Diot, 2007) observed that PCA is the linear transform which retains the highest variance among all transforms reducing the data to a certain number of dimensions, for any choice for the number of dimensions. They used this property to argue that PCA is in some sense optimal, for the task of anomaly detection.

PCA is based on *singular value decomposition*, and in some cases a singular value decomposition of data (interpreted as a matrix) can be applied directly without first computing the covariance of the data.

In this case, the use of SVD is merely an alternative approach for producing the same analysis of the data into features. Singular value decomposition of a matrix is more general than PCA, which assumes the matrix being analysed is symmetric. On the other hand, the underlying optimality interpretation is still applicable, although in a generalised form, depending on the specific case.

The singular-value decompositions used in this paper all seek to identify *features* which summarise the traffic data as compactly as possible. This is a widely used methodology in many fields. Mathematically, a feature is a linear combination of the original data fields. Intuitively, a feature represents some recognisable characteristic of the data. For example, in data from photographs of faces, "the nose" might be a feature discovered in the data in this way. The intuitive characteristics corresponding to the features discovered in the traffic data have not been identified in the literature applying SVD in this way (Lakhina et al., 2004), or by us, and it remains to be seen whether these intuitive characteristics will be discovered.

Antraff software

In order to carry out the analysis methods described above, and other methods, a software package called *Antraff* (Addie, 2016) has been developed. This software collects statistics of three different sorts:

(i) Byte frequencies of origins, destinations, or O-D pairs. See Figure 1.

(ii) Accumulation of matrices of end-to-end traffic, followed by analysis of these matrices, by singular-value decomposition. There are three different ways in which traffic matrices can be accumulated, which can be intuitively described as mean, variance, and covariance matrices.

(iii) Extraction of *eigenflows* associated with one of the three possible singular-value decompositions from the

http://doi.org/10.18080/ajtde.v5n2.106

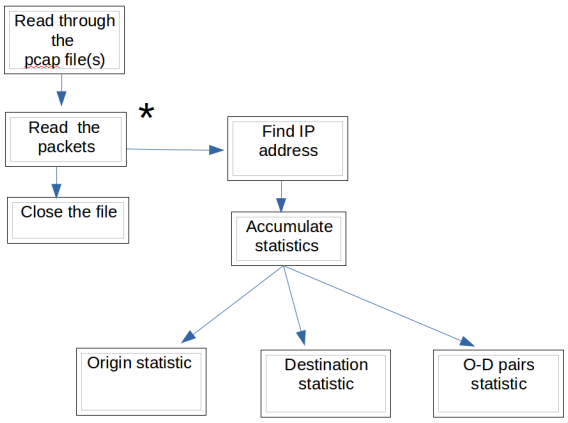


Figure 1: Algorithm for collecting origin/destination/O-D pair byte frequencies

previous analysis method.

Volume statistics

The total number of origins, destinations, or O-D pairs in the pcap file or collection of pcap files being analysed may be more than we are able to store information about. However, it will be satisfactory if the only sources, destinations, or O-D pairs about which information is not collected are ones for which the total bytes, associated with this origin, destination, or pair, is small. By storing pointers to origins, destinations, or O-D pairs in a heap which is ordered so that the entity with the least total bytes, over an interval of time exceeding a certain threshold, can be easily removed, it is possible to identify origins, destinations, or O-D pairs which can be removed from the list of those about whom statistics are being collected without having to collect the information about this object, except for short periods of time. The volume information collected in this way is stored in such a way that we can estimate the distribution of bytes associated with each origin, destination, or O-D pair.

Eigenvalue statistics

The origin and destination statistics described in the preceding sub-subsection are used to identify the *most important* (in terms of numbers of bytes) origins, destination, and O-D pairs. Three types of traffic matrices are then estimated:

(i) mean bytes, origin to destination;

(ii) $E(X^2)$ and variance of bytes, origin to destination;

http://doi.org/10.18080/ajtde.v5n2.106

(iii) covariance of O-D flows.

Note, the first two matrices are indexed by origin and destination IP addresses (not the actual IP addresses, but their rank), and the last of type is indexed by OD pairs (i.e. the rank of the OD pairs). Nevertheless, even in the last case, since each OD pair obviously also refers to an origin and a destination, the implications of all three matrices ultimately refer to a matrix which is indexed by origins and destinations.

Eigenflows

Once the singular decomposition of the traffic matrices outlined in the previous subsection has been completed it is possible to extract, from the traffic data, the *eigenflows* corresponding to each eigenvalue, and also the component O-D flows of each eigenflow. Each eigenflow is a linear-combination of O-D flows.

It is conceivable that there are highly important correlations between one O-D flow and another. For example, if a certain group of users engage in communication behaviour simultaneously, as a consequence of some shared activity of goal, this will lead to correlated O-D flows. The Antraff software is currently able to undertake singular value decomposition of traffic matrices and also extract both the eigenflows, and to display the O-D flows from which they are composed. However, it remains possible, at this stage, that the eigenflows are actually artificial, i.e. they are not actually a consequence of any genuine correlation between O-D flows.

Experiments

Source and Destination Byte Frequencies

The traffic associated with individual IP addresses exhibits the Pareto principle, which is that a small proportion of IP addresses is associated with a high proportion of the traffic. In other words, the distribution of bytes per destination/source is *heavy tailed*. This is illustrated in Figures 2 and 3.

On a log-log scale, these plots are approximately linear. Both curves exhibit a tendency to "roll off", i.e. the curve falls away from the linear shape to the right. If the population was such that this plot was perfectly linear on a log-log scale, in a theoretical sense for the whole population, estimates of the curve from a finite sample will always exhibit this type of roll-off because of sampling error.

These Figures *suggest* that most bytes are accounted for by the larger sources of traffic. However we can't actually read from these plots the proportion of bytes accounted by any initial portion (in decreasing weight) of sources. Figure 5 and 6 are designed to allow this. The Figures present the same information, but now the

http://doi.org/10.18080/ajtde.v5n2.106

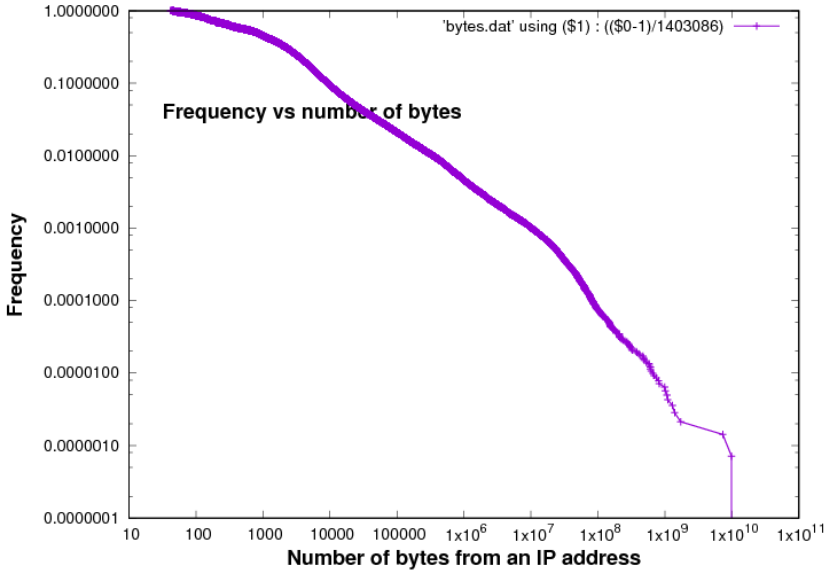


Figure 2: Bytes per source [Chicago 2015 data-set]

x-axis is the cumulative proportion of source and destination IP addresses respectively, in the data set, in order of weight, and the y-axis is the cumulative proportion of total bytes accounted for, by these IP addresses.

These Figures confirm the intuitive idea already introduced, that a small proportion of source IP addresses accounts for a large proportion of traffic, in bytes. However, the story is not *extreme*. To account for 90% of the bytes it is necessary to include at least 1% of sources, destinations, or O-D pairs, which is still quite a large number – approximately 2,000 in for the traces used in Figures 5, 6 and 7.

O-D Pairs

Another way to analyse the traffic is shown in Figure 7. This time, instead of focussing on which *sources* or which *destinations* account for the traffic, it is the *O-D pairs*. We are looking here for an understanding of the structure of traffic from the point of view of O-D pairs which may prove to be very important in the development of an end-to-end traffic model.

The statistical results of O-D pairs, in this paper, are different from other studies in regard of the number of IP addresses which are analysed. The authors of (Lakhina et al., 2004) have analysed the structure of Origin-Destination traffic in a trace of Internet backbone traffic by using PCA to systematically decompose the structure of OD flow time series. They describe the trace the analyse as including hundreds of origins and

http://doi.org/10.18080/ajtde.v5n2.106

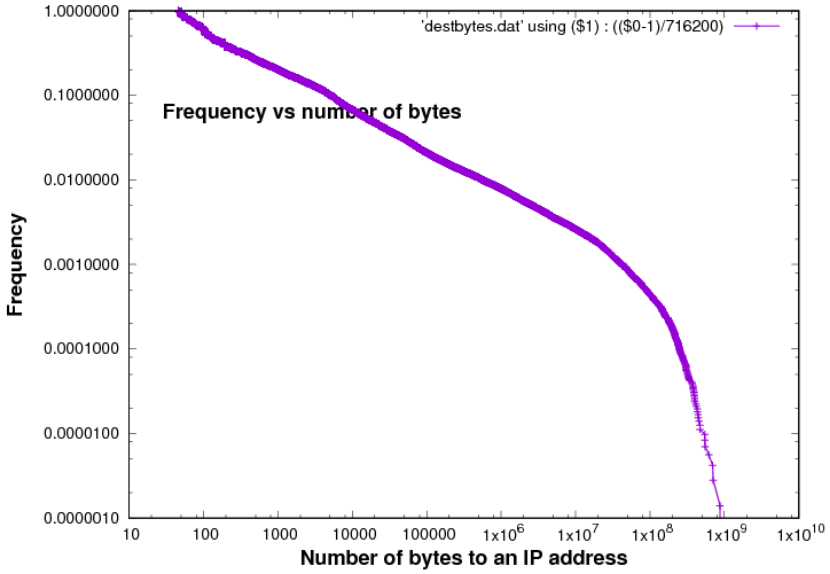


Figure 3: Bytes per destination [Chicago 2015 data-set]

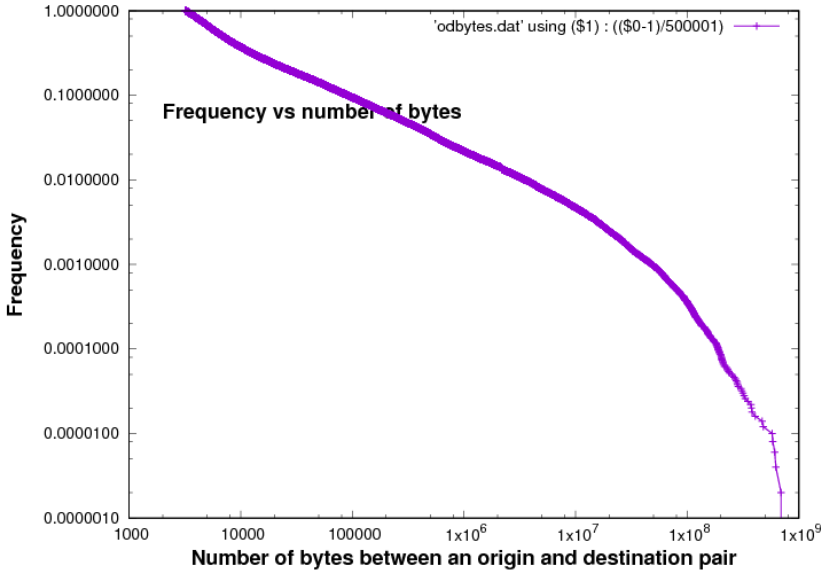


Figure 4: Bytes per O-D pairs [Chicago 2015 data-set]

http://doi.org/10.18080/ajtde.v5n2.106

Table 1: CAIDA data-sets (each data set consists of 10 data files, as in Date/Time column, each file has same prefix and same suffix)

Data-set Name	**Prefix**	**Date/Time**	**Suffix**
Chicago 2014	equinix-chicago-.dirA-2014	0320-130000, 0320-130100, 0320-130200, 0320-130300, 0320-130400, 0619-130000, 0619-130100, 0619-130200, 0619-130300, 0619-130400	UTC-.anon-.pcap
San Jose 2014	equinix-Sanjose-.dirA-2014	0320-125905, 0320-130100, 0320-130200, 0320-130300, 0320-130400, 0619-130000, 0619-130100, 0619-130200, 0619-130300, 0619-130400	UTC-.anon-.pcap
Chicago 2015	equinix-chicago-.dirA-2015	0219-125911, 0219-130000, 0219-130100, 0219-130200, 0219-130400, 0917-125911, 0917-130000, 0917-130100, 0917-130200, 0917-130300	UTC-.anon-.pcap
Chicago 2016	equinix-chicago-.dirA-2016	0121-125911, 0121-130000, 0121-130100, 0121-130200, 0121-130300, 0121-125911, 0317-130000, 0317-130100, 0317-130200, 0317-130300	UTC-.anon-.pcap

http://doi.org/10.18080/ajtde.v5n2.106

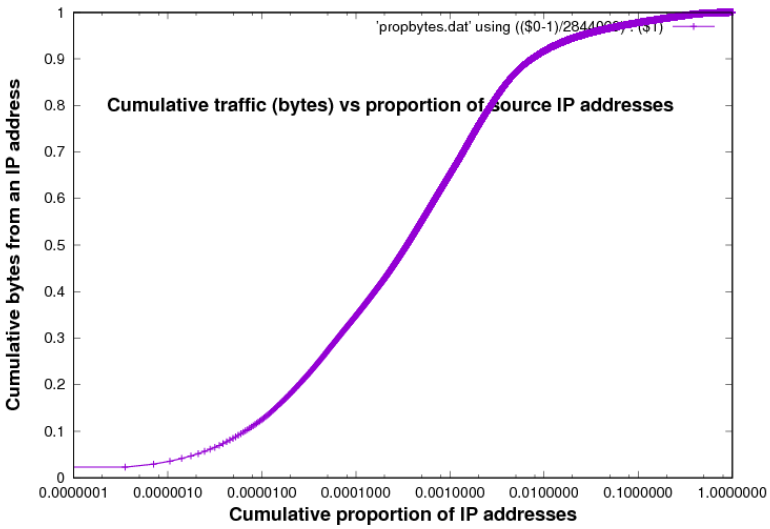


Figure 5: Cumulative sources bytes from an IP address vs Cumulative proportion of IP addresses [trace used: Chicago 2016 data-set]

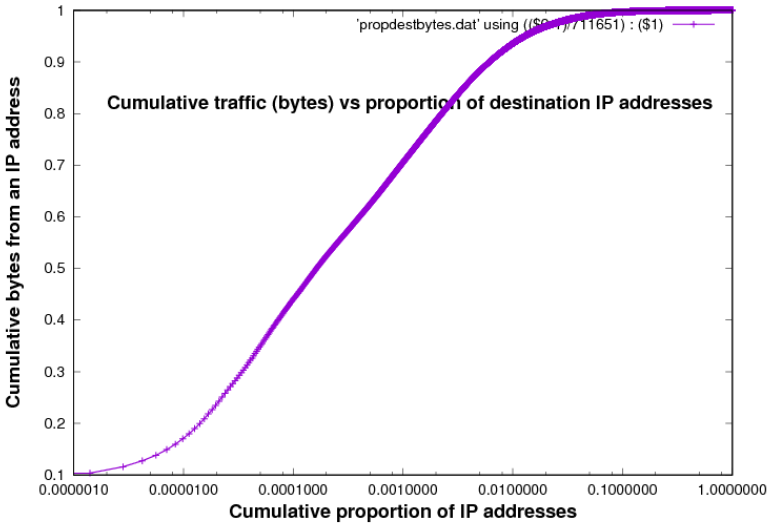


Figure 6: Cumulative destination bytes to an IP address vs Cumulative proportion of IP addresses [trace used: Chicago 2015 data-set]

http://doi.org/10.18080/ajtde.v5n2.106

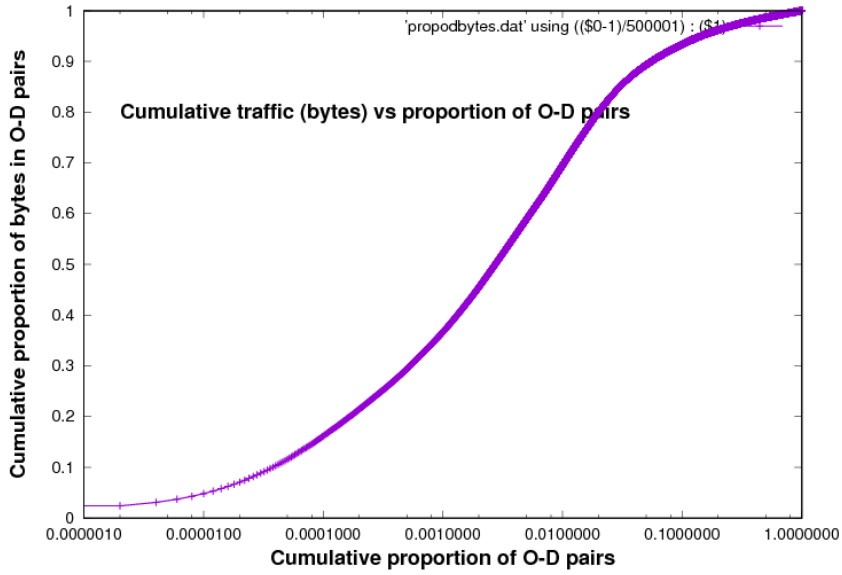


Figure 7: Cumulative bytes in O-D pairs address vs Cumulative proportion of O-D pairs [Chicago 2015 dataset]

destinations, and hundreds of flows. Another group of authors in (Susitaival, Juva, Peuhkuri, & Aalto, 2006) have analysed 1000 IP addresses to study the traffic characteristics of O-D pairs components, whereas the ones considered in this paper includes hundreds of thousands of origins and destinations, and millions of flows. Another difference is that the traces considered in this paper have durations of approximately one minute, whereas in (Lakhina et al., 2004) they consider traces lasting hours, days, or weeks. Figure 7 illustrates the relation between cumulative bytes in O-D pairs vs cumulative proportion of O-D pairs starting with the largest flows because of the importance of these flows. This Figure shows that the first 10% of O-D pairs covers about 90% of bytes.

O-D-pairs vs origins and destinations

Figure 8 show the number of O-D pairs found in several files of different traffic traces a certain point in the file, *against* the total number of sources and destinations IP addresses. The traffic traces are captured from Chicago and San Jose on 2014, 2015 and 2016 in US. This has been plotted using log scales for both axes. The curve is approximately linear, with a slope close to 1. In fact the slope, in general on the log-log scale, means how broad is the community and it measures the diversity. This means that the value of the slope measures the breadth of the community whose communication is taking place in the trace file.

The closeness to 1 indicates the degree to which communication in this trace is within closed communities.

http://doi.org/10.18080/ajtde.v5n2.106

In other words, the closeness of the slope in these Figures to 1 means that the individuals represented in the trace tend to communicate with the same relatively small group of friends.

Experimental results of the plots slope

The slopes in Figure 8 which are represent the relation between the number of sources and the number of destinations in x-axes and the number of O-D pairs in y-axes, are shown in Table 2.

If everyone communicated with only one other individual, the slope would be exactly 1, while if everyone communicated with everyone else, the slope would be 2. The slope will also be very close to 1 if there is a small number of central nodes, which we might call the gatekeepers, and every other individual communicates only with these gatekeepers. The data files, which are downloaded from CAIDA, in Table 2 come from different years, 2014, 2015 and 2016, and different places, Chicago and San Jose in the US.

Given the prominence of key sites in the Internet, the likes of Google, Youtube, and Facebook, it is therefore not surprising that the slopes in these Figures are close to 1 as shown in Table 2.

To investigate further the significance of the slopes, a t-test was carried out to compare one group against another and the results of these tests are as follows:

(i) the slopes of the source plots are less than those of the destination plots;

(ii) the four source slopes are different from each other;

(iii) the destination slopes in 2014-2015 are not significantly different, but the destination slope in 2016 is significantly larger than those in earlier years.

All the t-tests were un-paired and tested equality vs inequality except for the comparisons of source slopes vs destination slopes and the comparison of the Chicago 2016 destination slope vs the other destination slopes. The test of destination slopes vs source slopes was paired and tested the hypothesis that the source slope is less than the destination slope vs the alternative that it is not less. The test of the Chicago destination slope vs the other destination slopes was unpaired and tested the hypothesis that the Chicago 2016 slope is greater than the other source slopes.

Eigenflows

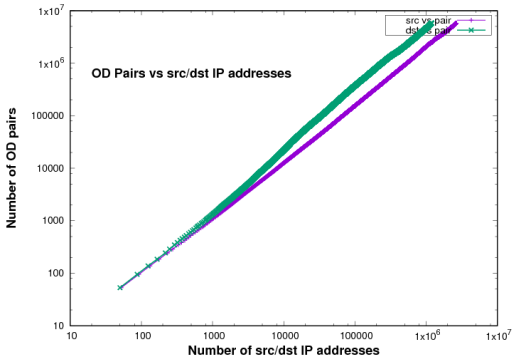
Figure 9 shows the eigenvalues of the three different types of traffic matrix which were investigated. It is noteworthy how similar the three types are. The covariance matrix eigenvalues are generally larger than those

http://doi.org/10.18080/ajtde.v5n2.106

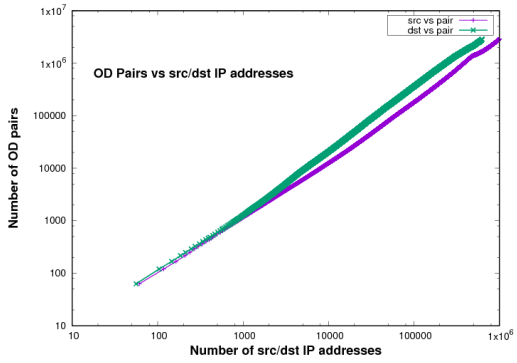
Table 2: Source and destination slope values of traffic traces

Data-set Name	File Name	Slope–S	Slope–D
Chicago 2014	equinix-chicago.dirA.20140320-130000.UTC.anon.pcap	1.19825	1.240841
	equinix-chicago.dirA.20140320-130100.UTC.anon.pcap	1.206898	1.240961
	equinix-chicago.dirA.20140320-130200.UTC.anon.pcap	1.196147	1.231688
	equinix-chicago.dirA.20140320-130300.UTC.anon.pcap	1.212603	1.23856
	equinix-chicago.dirA.20140320-130400.UTC.anon.pcap	1.205314	1.245019
	equinix-chicago.dirA.20140619-130000.UTC.anon.pcap	1.151398	1.156585
	equinix-chicago.dirA.20140619-130100.UTC.anon.pcap	1.159662	1.152868
	equinix-chicago.dirA.20140619-130200.UTC.anon.pcap	1.146592	1.194499
	equinix-chicago.dirA.20140619-130300.UTC.anon.pcap	1.198673	1.178119
	equinix-chicago.dirA.20140619-130400.UTC.anon.pcap	1.169297	1.236181
San Jose 2014	equinix-sanjose.dirA.20140320-125905.UTC.anon.pcap	1.11139	1.178774
	equinix-sanjose.dirA.20140320-130100.UTC.anon.pcap	1.117446	1.154508
	equinix-sanjose.dirA.20140320-130200.UTC.anon.pcap	1.111833	1.154293
	equinix-sanjose.dirA.20140320-130300.UTC.anon.pcap	1.108536	1.171695
	equinix-sanjose.dirA.20140320-130400.UTC.anon.pcap	1.119749	1.16328
	equinix-sanjose.dirA.20140619-130000.UTC.anon.pcap	1.119749	1.216095
	equinix-sanjose.dirA.20140619-130100.UTC.anon.pcap	1.091031	1.211002
	equinix-sanjose.dirA.20140619-130200.UTC.anon.pcap	1.088507	1.207736
	equinix-sanjose.dirA.20140619-130300.UTC.anon.pcap	1.100227	1.221282
	equinix-sanjose.dirA.20140619-130400.UTC.anon.pcap	1.094473	1.222404
Chicago 2015	equinix-chicago.dirA.20150219-125911.UTC.anon.pcap	1.156123	1.22055
	equinix-chicago.dirA.20150219-130000.UTC.anon.pcap	1.153414	1.242492
	equinix-chicago.dirA.20150219-130100.UTC.anon.pcap	1.12059	1.31508
	equinix-chicago.dirA.20150219-130200.UTC.anon.pcap	1.161377	1.234627
	equinix-chicago.dirA.20150219-130400.UTC.anon.pcap	1.160145	1.240479
	equinix-chicago.dirA.20150917-125911.UTC.anon.pcap	1.113016	1.079916
	equinix-chicago.dirA.20150917-130000.UTC.anon.pcap	1.113626	1.06874
	equinix-chicago.dirA.20150917-130100.UTC.anon.pcap	1.114851	1.067286
	equinix-chicago.dirA.20150917-130200.UTC.anon.pcap	1.119336	1.067437
	equinix-chicago.dirA.20150917-130300.UTC.anon.pcap	1.117681	1.085924
Chicago 2016	equinix-chicago.dirA.20160121-125911.UTC.anon.pcap	1.047565	1.242566
	equinix-chicago.dirA.20160121-130000.UTC.anon.pcap	1.064275	1.236514
	equinix-chicago.dirA.20160121-130100.UTC.anon.pcap	1.062016	1.234313
	equinix-chicago.dirA.20160121-130200.UTC.anon.pcap	1.063316	1.22516
	equinix-chicago.dirA.20160121-130300.UTC.anon.pcap	1.061707	1.237587
	equinix-chicago.dirA.20160317-125911.UTC.anon.pcap	1.088713	1.479135
	equinix-chicago.dirA.20160317-130000.UTC.anon.pcap	1.098664	1.474604
	equinix-chicago.dirA.20160317-130100.UTC.anon.pcap	1.099633	1.458606
	equinix-chicago.dirA.20160317-130200.UTC.anon.pcap	1.095182	1.482734
	equinix-chicago.dirA.20160317-130300.UTC.anon.pcap	1.095915	1.48362

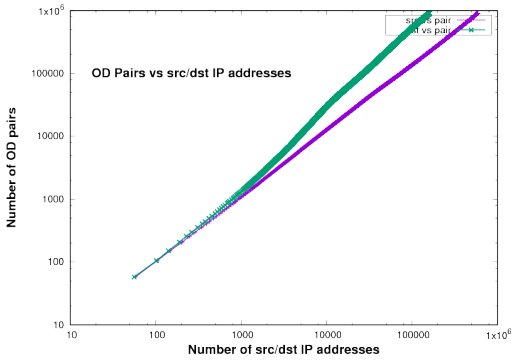
http://doi.org/10.18080/ajtde.v5n2.106



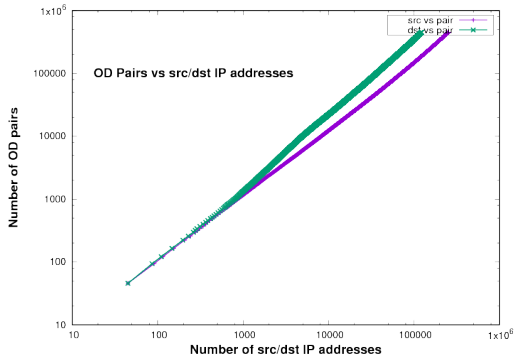
(a) San Jose 2014 data-set



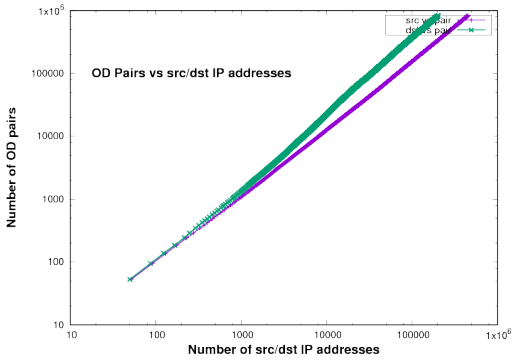
(b) Chicago 2014 data-set



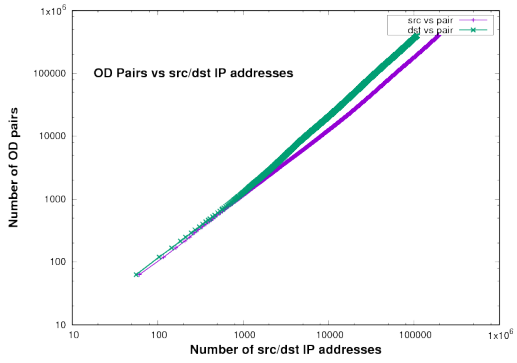
(c) Chicago 2016 single file



(d) Chicago 2015 single file



(e) San Jose 2014 single file



(f) community1.png

Figure 8: Community:Number of O-D pairs vs. Number of source and destination IP addresses for different data files[see Table 1]

http://doi.org/10.18080/ajtde.v5n2.106

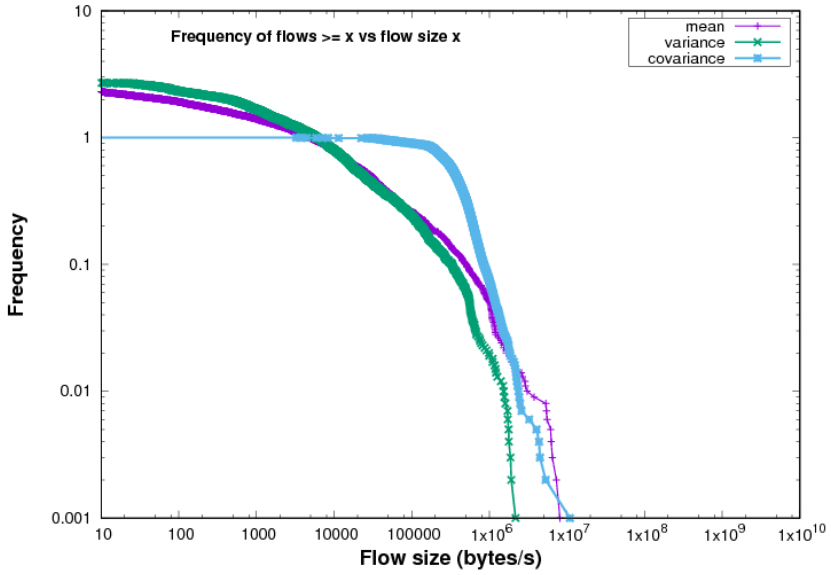


Figure 9: Eigenvalues of the mean, variance and covariance traffic matrices [trace used: Chicago 2016 data-set] from the mean and variance matrices, which indicates that the covariance matrix analysis is more effective in finding key features.

Figure 10 shows the *eigenflows* corresponding to the largest three eigenvalues of each type of traffic matrix. If there is significant correlated (in time) behaviour by users, these analysis methods should find it. However, the method does not in itself confirm that such correlations exist to a significant degree.

O-D Flow sizes

It has been suggested previously that flow sizes in the traditional sense of the number of bytes in an individual TCP connection have a power-law distribution (Crovella & Bestavros, 1997). Many papers have been written analysing traffic models based on this assumption, e.g. (Tsybakov & Georganas, 1998). However, in this paper the term *flow* has a different interpretation: it refers to all the traffic between a certain originating user and another destination user. The power-law character of the sizes of *this* type of flow does not necessarily follow from the power-law character of flows of the type considered in (Crovella & Bestavros, 1997). Nevertheless, the broadly linear shape of the curves in Figures 5–7 supports the power-law character of flows in the sense this term is used here.

Figure 11 is the same as Figure 7 except that the vertical scale has been transformed in the same manner as a quantile plot, of the type used to check normality of a distribution. The straightness of the plot in Figure 11

http://doi.org/10.18080/ajtde.v5n2.106

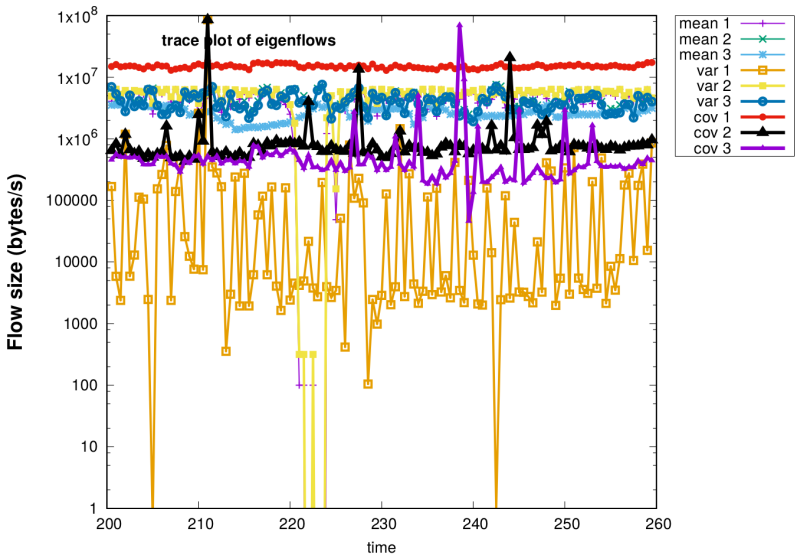


Figure 10: Eigenflows corresponding to the first three eigenvalues of the singular value decomposition of the mean, variance, and covariance traffic matrices [Chicago 2016 data-set]

indicates that the rank of an O-D pair selected by drawing a byte randomly, from all the bytes passing through a link, and then finding the O-D pair between which it was being transmitted, is approximately log-normally distributed. Note: the smallest 15% of flows are not included in this plot.

Conclusion

A Pareto principle applies to every aspect of the traffic. Most users contribute very few bytes, and most bytes are from a very small proportion of users. This applies to sources, destinations, and O-D pairs. The degree to which this is the case can be measured. It is the slope in the curves shown in Figures 2–4. These parameters are characteristic of the underlying social network (without in any way referencing any single individual or flow) which we can measure, and investigate. Are they the same in different places, and at different times? Can we use these characteristics to better understand how to serve the needs of the communities using our networks?

To convey the quality of power law distributions we have frequently emphasised the extreme values and this may give a false impression. It is not only the fact that a high proportion of bytes are accounted for by a small number of flows, while another (different) high proportion of flows account for a very small proportion of bytes, which is revealed in Figures 5–7. The power-law character applies across the full range of flow sizes and this is likely to be important in order to be able to apply it effectively.

Also, we see in Figure 8, a slightly different characterization of social behaviour is revealed. These features measure the degree to which all communication is mediated through a limited number of gatekeepers.

http://doi.org/10.18080/ajtde.v5n2.106

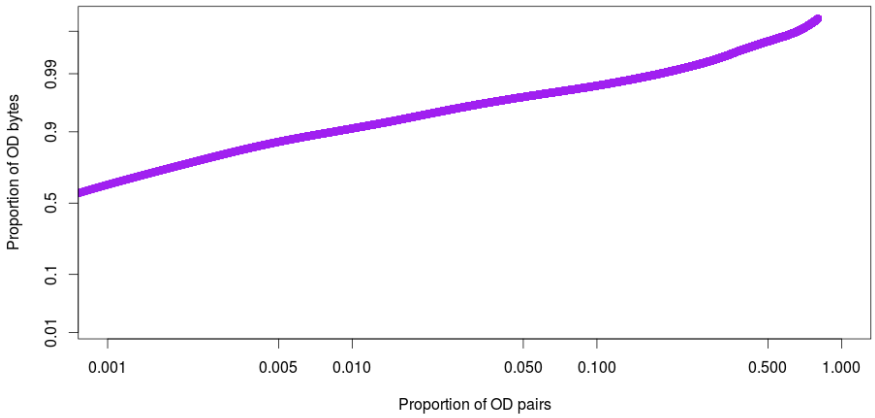


Figure 11: Log-normal character of OD flow sizes [Chicago 2015 data-set]

Lastly, singular value decomposition of traffic matrices of several different types was used to identify *eigenflows* of different types. It has been suggested (Lakhina et al., 2004) that these eigenflows may reveal important features of network traffic. However, it remains possible that eigenflows are merely artefacts, i.e. the features they emphasise are more accidental than characteristic of the underlying social network. It remains to determine the statistical significance of the eigenvalues of the singular value decomposition of traffic matrices and therefore of the eigenflows. Methods for determining this statistical significance are under development as part of the research reported here. If some or all eigenflows are truly statistically significant they may provide a very important insight into underlying social behaviour.

These are probably not the only ways to characterise social behaviour. However, the statistical analyses presented in this paper, and in particular the diagrams generated from them, clearly show that there are important characteristics of the underlying social network which can be identified from traffic data.

References

Adas, A. (1997, Jul). Traffic models in broadband networks. *Communications Magazine, IEEE*, *35*(7), 82-89. doi: 10.1109/35.601746

Addie, R. G. (2016). *Antraff traffic analysis software user manual* (Tech. Rep.). USQ.

The CAIDA UCSD Anonymized Internet Traces 2014 - [20140320]. (n.d.). Retrieved from http://www.caida.org/data/passive/passive_2014_dataset.xml

Center for applied internet data analysis. (2016). (http://www.caida.org)

http://doi.org/10.18080/ajtde.v5n2.106

Chandrasekaran, B. (2009). Survey of network traffic models. *Waschington University in St. Louis CSE*, *567*.

Crovella, M., & Bestavros, A. (1997). Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, *5*(6), 835–846.

Erramill, V., Crovella, M., & Taft, N. (2006). An independent-connection model for traffic matrices. In *Proceedings of the 6th acm sigcomm conference on internet measurement* (pp. 251–256).

Fan, J., Xu, J., Ammar, M. H., & Moon, S. B. (2004). Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme. *Computer Networks*, *46*(2), 253–272.

Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E. D., & Taft, N. (2004). Structural analysis of network traffic flows. In *Acm sigmetrics performance evaluation review* (Vol. 32, pp. 61–72).

Ringberg, H., Soule, A., Rexford, J., & Diot, C. (2007). Sensitivity of PCA for traffic anomaly detection. *ACM SIGMETRICS Performance Evaluation Review*, *35*(1), 109–120.

Susitaival, R., Juva, I., Peuhkuri, M., & Aalto, S. (2006). Characteristics of origin-destination pair traffic in funet. *Telecommunication Systems*, *33*(1-3), 67–88.

Tsybakov, B., & Georganas, N. D. (1998, September). Self-similar processes in communications networks. *IEEE Transactions on Information Theory*, *44*(5), 1713–1725.

Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, *91*(433), 365-377.

http://doi.org/10.18080/ajtde.v5n2.106

Making ICT Decommissioning Sexy!

Challenges and Opportunities

Peter Hormann
Melbourne Networked Society Institute, University of Melbourne

Abstract: The author contends that the ICT industry has become good at accumulating large inventories of ICT systems without paying sufficient attention to the associated technical debt and cost. He has experienced organisations that have repeatedly failed to retire old systems, are struggling to modernise, are burdened by complexity and can't change at a pace demanded by their internal stakeholders and their customers. Why is this case and why isn't the decision to switch off ageing, low business value, poor strategic fit and high-risk systems an easy and obvious business case? This paper explores the challenges of decommissioning decision making and uses a framework and case study analysis as a means to improve the timeliness and financial motivations. It should be of interest to ICT and business executives that struggle with balancing a tight budget around existing systems, with the need to invest in systems modernisation to maintain organisation competitiveness and productivity.

Introduction

"Decommissioning is never sexy and often struggles to get funded."

– A former CFO of a major Australian ICT company

Based on the author's industry experience, decommissioning is often not seen as a fundable business priority at the highest levels of an organisation. He has found that the timely decommissioning of ICT systems, incorporating applications, platforms, networks and devices, can be a perennial and frustrating challenge for technology managers. More so, the accumulation of legacy ICT systems inhibits organisational modernisation, innovation, agility and the speed to market demanded by new revenue opportunities. Ironically the bias for capital funding of new revenue opportunities can starve an organisation of the funds needed to decommission, rationalise and ultimately modernise its ICT systems. Decommissioning can be treated as an operational overhead and relegated a low priority until other factors (e.g. risk of failure, legal liability, real estate constraints, power and supportability) cannot be ignored.

http://doi.org/10.18080/ajtde.v5n2.109

The 'true cost' of maintaining a system of low business value, high risk/cost of failure and poor strategic fit can be underestimated. It often neglects the real cost of ICT technology debt, architectural complexity, real-estate value, and environmental impact (e.g. energy usage and associated carbon emissions). Unlike discretionary new investments, the decision to decommission is not a question of whether to invest or not, but rather discretionary only as a matter of timing. Eventually, all systems need to be decommissioned, and in the meantime the direct and potentially hidden costs to maintain legacy systems are an ongoing business overhead.

The business case for decommissioning a system can be nontrivial and its form can vary from organisation to organisation. Making decisions focussed on **Total Cost of Ownership** can be overly simplistic and mask other less apparent business factors and opportunities. According to Murphy, P. (2011), decommissioning has some unique business requirements and there is a need to look beyond just cost-benefit analysis as this is too narrow a viewpoint. On this basis, this paper takes a more systemic approach to decommissioning decision making by incorporating systems' assessment of **Business Value**, **Strategic Fit**, **Ease of Decommissioning** and **Risk Manageability**.

The paper starts with a review of the opportunities and the challenges for decommissioning. It then introduces a framework for assessment and a case study use of the framework undertaken in partnership with the City of Melbourne Council, Victoria, Australia. The paper concludes with 10 key recommendations for improving decommissioning outcomes that ultimately could improve corporate profitability, delivery and sustainability.

Scope

There are many publications covering systems' lifecycle management. However, in the author's experience, the complexity of decommissioning decision making is often underestimated. A generalised technology lifecycle from an ICT user's perspective is shown in Figure 1. As highlighted, this paper focuses on the Decommissioning Decision phase in the systems' lifecycle.

The Opportunities

A traditional cost-benefit business case for decommissioning will be comprised of the systems' apparent cost of ownership versus the costs to decommission either in the systems' entirety or as part of a functional migration to an alternative system. The usual cost of ownership, so-called "hard" costs, include licensing, support, maintenance, infrastructure and associated support staff. For an older system, "hard" costs may already be small as previous savings efforts have reduced these costs to an essential minimum. A business case

http://doi.org/10.18080/ajtde.v5n2.109

determined on "hard" costs alone is often found not to provide a worthwhile return on investment and therefore decommissioning can repeatedly be forestalled.

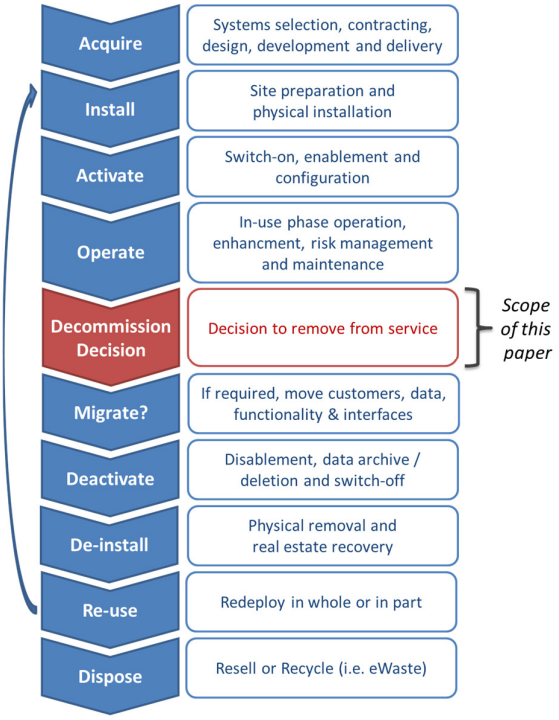


Figure 1: Systems lifecycle and the scope of this paper

A prudent corporate business case necessitates justification based on the financially "hard" costs, however, there can be other financially "soft" and "intangible" value. The "soft" and "intangible" value are potentially intuitively compelling but often based upon subjective justifications only. This additional value should nonetheless be of business interest and can provide additional business justification over hard cost savings alone; see Figure 2.

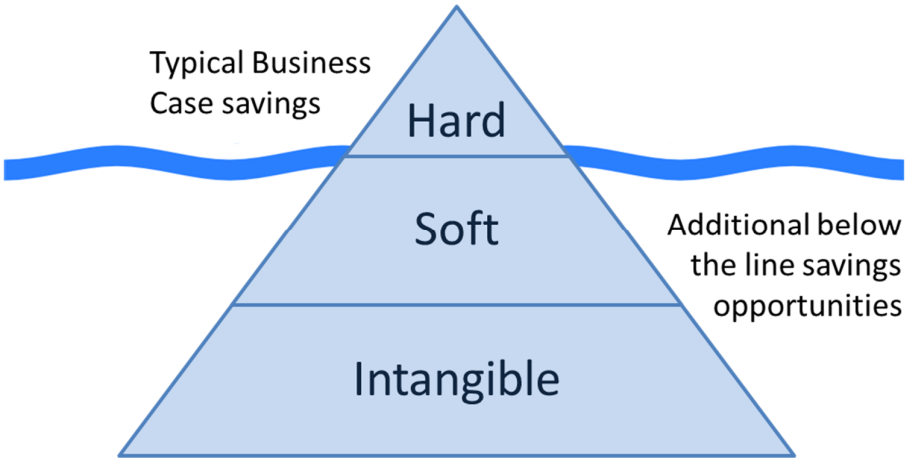


Figure 2: The decommissioning iceberg: above and below the line business savings

http://doi.org/10.18080/ajtde.v5n2.109

Table 1: Example hard, soft and intangible decommissioning value

Example Decommissioning Value		
Hard	Soft	Intangible
• Reduced applications licensing, contract support and maintenance. • Reduced compute, data storage / backup, networking and spares. • Reduced specialist staff, skills maintenance and associated training. • Reduced operational costs including alarm management and truck rolls. • Reduction in development and test platforms costs. • Reduced datacentre real-estate, energy and cooling costs. • Scrap equipment value.	• Avoided asset and lifecycle replacement costs. • Reduced legal and regulatory responsibilities – eg. privacy, security, data retention. • Reduced backup, business continuity and disaster recovery complexity. • Physical to virtualised platform consolidation benefits. • Environmental savings (eg. carbon emissions from fossil fuelled electricity). • Improved corporate sustainability through reduction in low-value producing expenditure and waste.	• Improved profitability. • Consolidated business processes and simplified user experience. • Modernised ICT fleet enabling more cost effective and faster time to market. • Improved systems' reliability and reduced risk. • Reduced architectural complexity and simplified technology roadmaps. • Simplified systems integration testing. • Reduced systems knowledge acquisition & preservation costs. • Reduced need for datacentre floor, power and cooling expansion.

Definitions for hard, soft and intangible savings:

- **Hard.** The financially tangible direct and bankable cost savings that could be recognised if a system was decommissioned, either in its entirety or functionally migrated to an alternative system. These savings will reduce an organisation's committed budget expenditure.
- **Soft.** The expenditure that is often uncommitted from a financial budget perspective, but in all likelihood could or should be spent to mitigate growing risks. These discretionary, but likely, future savings can be difficult to materially recognise in a financially focused business case.
- **Intangible.** The costs and upside value that is intuitively apparent, but buried amongst other shared systems costs, dependencies, processes and complexity. The lack of traceability and attributability of these costs makes them virtually impossible to include in a business case.

Examples of hard, soft and intangible savings value are shown in Table 1.

http://doi.org/10.18080/ajtde.v5n2.109

Figure 3: The challenges of ICT decommissioning

Challenges – Why can it be so difficult?

This paper is less concerned about the idle server platform running in the corner of a data centre or under an employee's desk; the rationale to identify, migrate, and switch-off any such systems should be relatively trivial, cheap and obvious. Instead, this paper is focussed on the ageing platforms that still have business dependencies making them technically, operationally, organisationally, commercially and legally difficult to be removed. There may also be customer dependencies to be managed. Addressing all these aspects in an effort to turn off ageing systems is challenging. Figure 3 lists many of the likely and associated considerations. This list is by no means exhaustive.

There are three other areas — data lifecycle management, financing and risk management — that warrant further discussion.

http://doi.org/10.18080/ajtde.v5n2.109

Data Lifecycle Management

Sound data lifecycle management is key to decommissioning ageing systems. Establishing exactly what data is located where, and how different systems process it, can require significant effort. Organisational data ownership can be unclear and in effect be shared across product, application, process, IT infrastructure, legal and customer groups. Without clear data ownership, organisations can be challenged by data preservation requirements, which may be mandated, perceived or unknown, and as a result, they find it difficult to know what data can be purged and what needs to be maintained in an archive. With no one group having clear ownership, an organisational default can become a practice of keeping everything "just in case".

Without clear data systems knowledge and ownership that can determine whether to purge, archive or migrate data over time, the best an IT infrastructure team can do is maintain and expand a system to avoid it failing or underperforming. The inadvertent demand to retain old data drives up applications support costs, processing infrastructure requirements, data backup window sizes, network capacity and software licensing costs. The challenges of data lifecycle management and the long term storage of data are further described by Lambrechts, J. (2014) and Hormann, P. (2014).

Financing

Finding a funding source for an ongoing decommissioning program can be a perennial challenge. Who should pay? Is it a capital or operational expenditure? What are the financial incentives and motivations? Who benefits from the financial rewards?

Capital Expenditure versus Operational Expenditure

Our system of business valuation and taxation generally promotes and rewards the use of capital expenditure to build new revenue earning investments. Funds to build a new or augment an existing system are usually classed as a capital expenditure. For example, capital expenditure can be used for systems infrastructure / software purchases, site preparation, development, delivery, installation and associated labour and training. The resulting system costs are depreciated from a taxation perspective over the expected lifetime of the asset. Importantly a system that has already been capitalised up front cannot be capitalised again at the end if its lifecycle. As a result, decommissioning activities are usually classed as an operating expense.

Operating expenditure is typically the recurring annual costs for running a system. These costs include staff, training, spares, licenses, maintenance and support agreements, and data centre hosting (e.g. real-estate, energy, cooling) charges. Operational expenditure is often scrutinised for waste and a company is rewarded when it is reduced. If an exercise in

http://doi.org/10.18080/ajtde.v5n2.109

systems decommissioning is difficult, and therefore costly, and the apparent operational cost savings are small over business case evaluation periods, then it is easy for a decommissioning project not to be funded. This is despite the inevitability that all systems must have an end, whether planned or as a result of a failure event.

As an exception, it may be possible to capitalise the funding for decommissioning where an old system needs to be removed to "make ready" for the building of the new; not unlike the demolishing of an old building as site preparation for the construction of a new building. From a CFO's perspective, it can be financially prudent to bundle the decommissioning of an old system with the capitalisation of its physical or functional replacement. However, the costs of the "make ready" can be seen as punitive to the business case of the new. This is especially the case where easy to access data centre space exists elsewhere and there is no reward or recognition to do otherwise. The efforts to work around the "make ready" costs can then drive a suboptimal outcome from a business and technical perspective that adds to the inadvertent sprawl of systems. An unfortunate consequence is that old systems will be left running and in place until their physical locations are required by a project that can't find more convenient space elsewhere.

Importantly, build budget overruns cannot be allowed to consume funds assigned for systems decommissioning. As such any funds, be they capital or operational, that are budgeted for decommissioning should be explicitly quarantined to avoid forestalling systems decommissioning in whole or in part to another time. This is especially the case where funding is strictly limited, the process for requesting new funds can be onerous and punitive, or decommissioning can be side-stepped through official or unofficial descoping practices.

Funding Sources

A business funding source for system decommissioning may not always be obvious as ongoing costs and risks are owned by one part of the organisation such as an IT department and the business need by another part such as a product manager. While a product manager may perceivably "need" a system while it is free of charge to them, this may not be the case if their product's profitability was tied to the system's real costs and risks. Alternatively, the benefits of decommissioning can be shared by many, but are not separately compelling enough for one part of an organisation (e.g. IT) to fund the opportunity on its own.

There are a number of potential ways to fund decommissioning initiatives; five example approaches with their pros and cons are summarised in Table 2. Others such as a sinking or accumulation fund may also be worth considering.

http://doi.org/10.18080/ajtde.v5n2.109

Table 2: Decommissioning funding sources – five key approaches

	Reactive	System Case-by-case	Funds Redirection	Funding Pool	Asset Retirement Obligation (ARO) Fee
Description	• Wait for a compelling event (eg. data centre move, corporate merger) or systems failure. • Wait for the last customer to exit.	• The deemed systems owner build's a decom business case for each systems decom opportunity.	• Reward a systems maintenance vendor to undertake decom projects on behalf of the business.	• Establish a dedicated funding pool for decom projects. • Grow the pool through centralising the savings.	• Build a retirement funding pool through an upfront fee on purchase, or an ongoing usage fee. • Fee can escalate over time.
Pros	• Allows problem to be ignored until it really matters to someone who cares enough to upgrade. • Enables the status quo to prevail until no one cares if the system is simply switched off.	• Treats decommissioning like any other investment activity and may not be treated as a business priority.	• Uses existing systems maintenance funds. • Disrupts the vendor's status quo with positive incentives. • Vendor no longer needs to retain and train staff in old technology.	• Ensures decom projects are considered on their own merits. • Centralises the benefits, expertise, effort and visibility.	• Enables a foreseeable end-event to be incorporated into the formative business case. • Ensures those that benefit from the system are also responsible for its end of life.
Cons	• Unexpected business disruption. • Unplanned expenditure. • More systems to manage. • Extra systems complexity makes new developments more costly.	• Lack of perceived urgency may mean projects are never funded. • Projects with low medium to longer term returns don't' get funding. • Small low return initiatives don't get funding.	• Financial benefits flow to the vendor first; financial savings to business takes much longer. • Difficult where multiple vendors involved. • Risks not shared.	• Easy to cut funding for the sake of short term corporate benefit. • Savings are centralised, rather than of benefit to the team that makes the savings.	• Difficult to upfront predict the ultimate cost to decom. • Perceived as an unnecessary internal business case "tax".

Worthy examples of raising funds for decommissioning (amongst other savings opportunities) are Microsoft's formative equipment recycling fee described by DiCaprio, T. (2012) and more recently Microsoft's internal carbon fee described by DiCaprio, T. (2015):

- **Recycling Fee** - introduced to cover the costs of recycling old electronic hardware as part of a policy for retired technology. The fee is collected at the time of equipment purchase and is centrally managed via an IT Asset Disposition program.
- **Carbon Fee** - helps raise awareness and drive accountability to enable a transparent and environmentally sustainable business model. It provides both an incentive and financial support for internal stakeholders to take responsibility for cutting energy costs

http://doi.org/10.18080/ajtde.v5n2.109

and reducing the organisation's carbon emissions through investment activities such as systems decommissioning.

Risks

Ageing systems can often be out of mind and out of sight, especially where much of the original costs have been progressively pruned and the system appears well down a corporation's long-tailed systems list. This is not to say that the system is not important, but if it is not a high-ranking core system (e.g. by total cost of ownership), then it is not likely to have a dedicated system owner and can be relegated a low priority from a management attention perspective.

A system without direct oversight and understanding can unknowingly become unsupported or unsupportable. Unwittingly systems costs such as licensing, support and maintenance, and other potential risk mitigating controls can be defunded without an essential understanding of the growing business risks.

Managing risk requires the understanding of the likely failure modes, the likelihood of failure, and the business consequences in the case of a failure. A system with a high potential to fail, or with a large impact on the business in case of failure, should be subject to effective risk controls that bring the system to an acceptable level of manageability. Risk controls could include managing system hardware and software, unplanned incident contingencies, preventative maintenance, availability and performance monitoring, system backups, documented disaster recovery processes, availability of spares, security patching, and logical/physical capacity planning. Risk controls should not be taken for granted, and their ongoing efficacy (including funding) should be regularly tested and evidential data collected.

An example risk assessment matrix based on Queensland CIO, (2016) is shown in Figure 4. It shows a generalised system without any controls having a risk likelihood of "4. Likely", a consequence of "4. Major" and an overall risk rating of "High" (C). With the addition of risk controls the overall risk is reduced to "Medium" (B), but is still short of business determined maximum acceptable "Low" risk rating (A).

If the costs of sufficient risk controls are operationally unaffordable and the maximum acceptable risk has not been achieved, then there is further justification for systems decommissioning where the cost of the risk controls are potential "soft" savings.

A decommissioning assessment framework

As part of this study, a detailed decommissioning assessment framework has been developed. The framework is focussed on gathering a wide range of both subjective and

http://doi.org/10.18080/ajtde.v5n2.109

objective measures and turning these into a quantifiable score across a range of business criteria. Its key components include:

1. **Questionnaire** – a list of 90 questions to be completed by the system owner.
2. **Question context** – further explanation about the question.
3. **Assessment scoring guide** – a 'how-to' for determining a score
   - 0 = poor (or unknown), to
   - 5 = good
4. **Business Criteria** – i.e. Business Value (BV), Strategic Fit (SF), Total Cost of Ownership (TCO), Risk Management (RM) and Ease of Decommissioning (EoD).
5. **Question assessment** – the numeric score assessment based on the questionnaire.
6. **Question score weighting** – a relative adjustment of importance between questions.
7. **Score summary table** – the final weighted and normalised results.

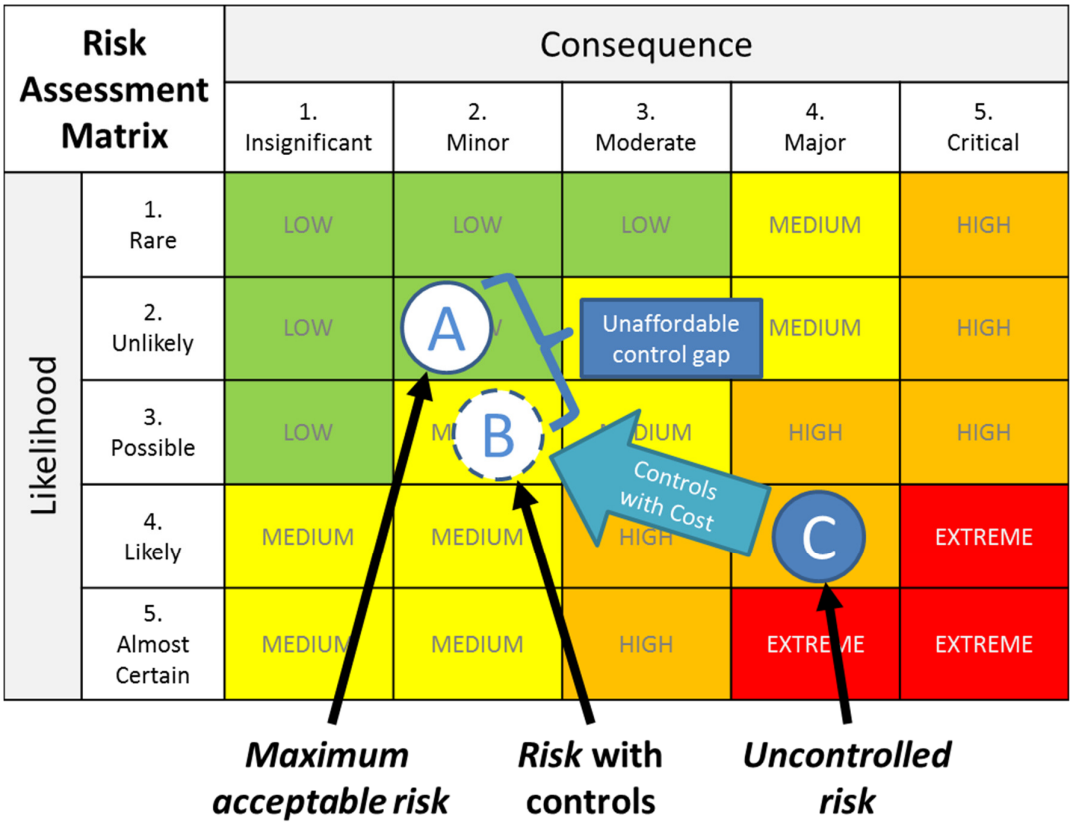


Figure 4: A Risk Assessment Matrix – example (based on Queensland CIO, 2016).

Figure 5 shows a blank example of the framework questionnaire, assessment and results table. The application of the framework is demonstrated in the following case study undertaken with the City of Melbourne Council. Readers are encouraged to generate their own assessment framework and are welcome to contact the author for further help.

http://doi.org/10.18080/ajtde.v5n2.109

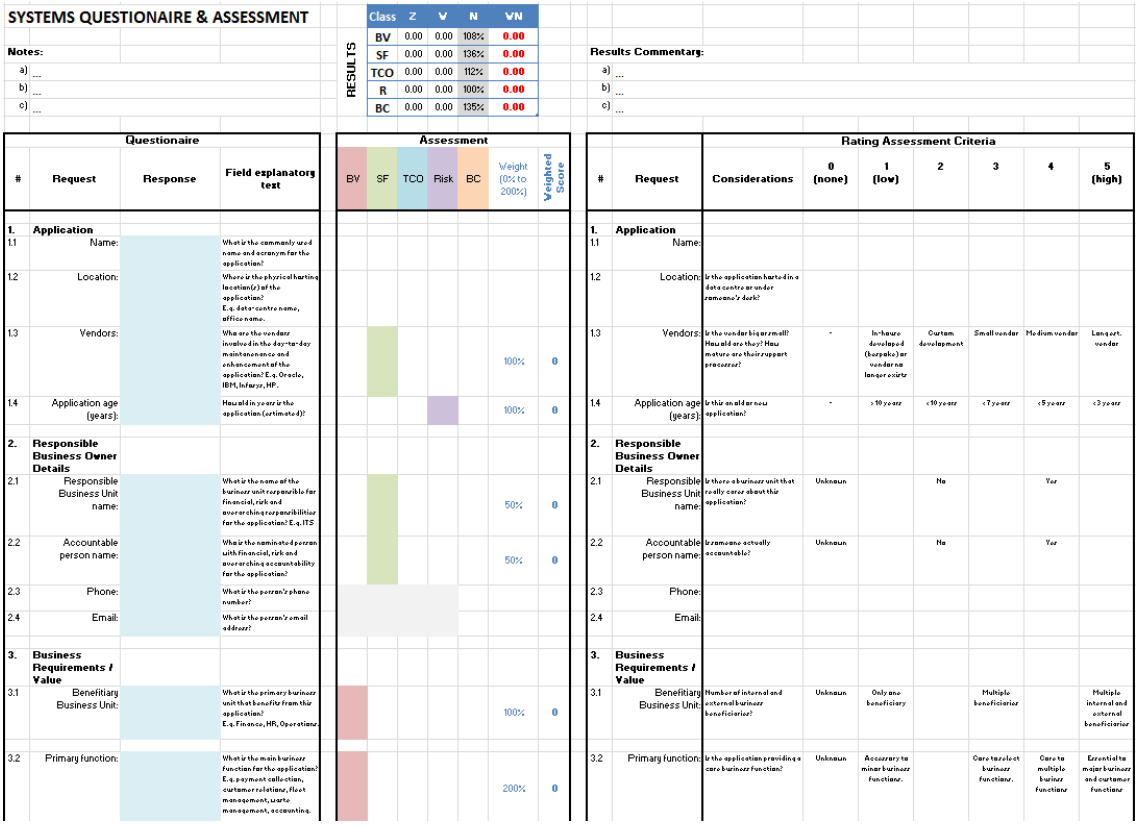


Figure 5: ICT Decommissioning Assessment Framework - sample

Case study - City of Melbourne

Background

The City of Melbourne IT systems architecture is comprised of about 100 key applications. There is also a long tail of other applications that are not actively managed by the IT team. Over recent years there has been a significant effort to rationalise and virtualise old systems with a high degree of success. There are, however, some old applications and associated platforms that remain running because of residual business process requirements and the need for ongoing data retention and accessibility. Like many councils, the City of Melbourne has mandated data preservation requirements; but it struggles with developing and implementing policies, processes and technologies to do this consistently well. By default, it is seen as easier and perceivably cheaper to continue to manage and sustain old systems.

In this study, four applications have been reviewed within the City of Melbourne's IT portfolio. Two applications, **PPR** and **CPC-Pro**, are 15 to 20 years old and considered as problematic decommissioning candidates. Another, **Vantive,** is of similar age but has recently been decommissioned after a successful business justification was completed.

http://doi.org/10.18080/ajtde.v5n2.109

Lastly, **MS-Exchange** is a modern platform and included as an assessment framework cross-check of a system that should intuitively remain current for the foreseeable future.

Methodology

The following steps were undertaken to complete the decommissioning assessment:

1. A questionnaire for each application completed by City of Melbourne system owners.
2. Assessment of the questionnaire responses by an independent assessor.
3. Applications / Systems which rate poorly against the business criteria for retention are ear-marked for its decommissioning.

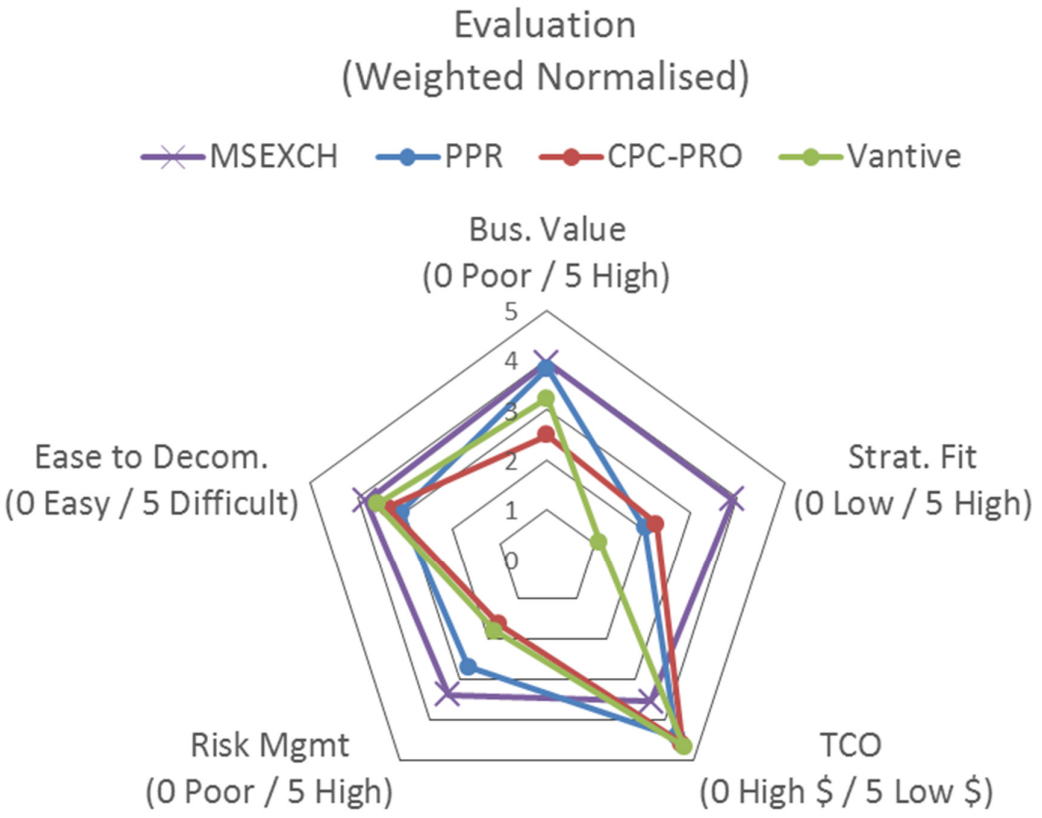


Figure 6: City of Melbourne decommissioning assessment results (radar chart)

Findings

The results of the analysis are shown in Figure 6 and Figure 7. System criteria scores closer to zero are indicators of positive decommissioning potential, while scores closer to five indicate a system that's worth retaining; at least for the time being.

A system of high value would rate all fives, while an obvious system candidate for decommissioning would be rated all ones. As can be seen from the results, none of the systems are easy or obvious decommissioning targets. A business case based on only TCO in

http://doi.org/10.18080/ajtde.v5n2.109

each case is likely to fail as there is little "hard" cost saving to be made and where the decommissioning exercise is likely to be medium to high in difficulty (and therefore high cost).

Observations about the results:

- **MS Exchange** is a relatively new and modern application and can be seen to rate highly in all categories. It offers good business value, is of good strategic fit, a reasonable TCO, is well risk managed, and could be reasonably easy to exit (e.g. migrate to a cloud-based email platform).
- **PPR**, **CPC-PRO** and **Vantive** all show low TCO (rated as 5 on the scale) where support, licensing and platform costs have been already minimised. They also have been evaluated with a medium level of difficulty for decommissioning, meaning the exercise requires some level of functional and data migration to a newer equivalent, and a fair degree of investment. All 3 are showing poor strategic fit, which is not unexpected given the systems' age of greater than fifteen years.
- **Vantive** particularly differs from PPR and CPC-PRO in regard to its lower strategic fit and increased risk management issues. As a result, and in the timeframe of writing this report, Vantive has been justifiably and successfully decommissioned.

As can be seen, the future of PPR and CPC-PRO are indeterminate. At a minimum, and if they are not to be decommissioned as yet, then investment is required to improve the risk management of these systems for the sake of minimising the likelihood of systems failure. The cost of the risk management may nonetheless make these systems decommissioning candidates in the more immediate term.

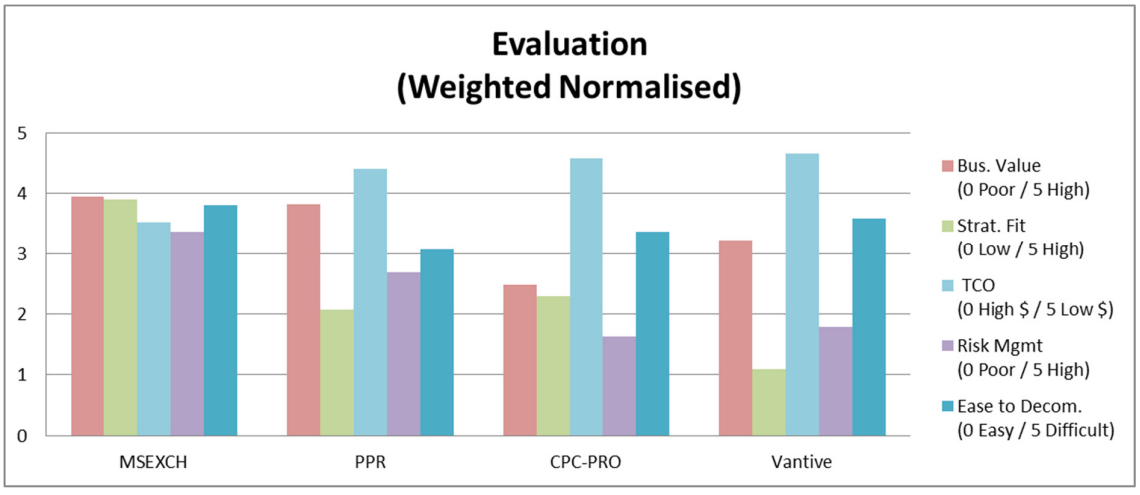


Figure 7: City of Melbourne decommissioning assessment results (bar chart)

http://doi.org/10.18080/ajtde.v5n2.109

Recommendations

Here are the authors 10 recommendations for decommissioning success:

1. **Establish a dedicated and expert team** whose prime responsibility is to switch off old systems. Head the team by a forthright, appropriately motivated and authoritative manager (a decommissioning Czar if you will).
2. **Identify an accountable executive** that has ultimate responsibility for systems risk management and cost (e.g. the CIO).
3. **Ensure there is a clear ownership of corporate data** with responsibilities for (i) purging low/no value data, (ii) preserving data that is important for the business, and (iii) meeting mandated data retention requirements.
4. **Establish a funding pool** from which to start and run an ongoing decommissioning programme, the savings from which help maintain and grow the pool.
5. **Create a decommissioning assessment framework** that evaluates financial savings, business value, strategic fit, risk management and other decommissioning benefits.
6. **Incorporate appropriate costs and subsidies** into new systems' business cases that encourages (i) migration and exiting of old systems, (ii) data centre space reuse and consolidation, (iii) good data lifecycle management, and (iv) the costs of inevitable future systems retirement obligations.
7. **Design for decommissioning** by ensuring new systems' builds are hardware independent, virtualisation/cloud friendly, modular and are designed to cost-effectively scale up and ultimately scale down.
8. **Quarantine project decommissioning funding** to ensure it cannot be subsumed by cost overruns of new systems build activity and/or descoped and delayed. Don't celebrate the completion of the new, until the old is also turned off.
9. **Apportion growing responsibility and costs** to those who resist systems decommissioning for specific functional, product or customer-specific reasons.
10. **Build decommissioning into architectural and capability roadmaps** so that systems end of life and strategic alternatives are clear and apparent. Avoid any new developments on systems with an impending end of life.

Conclusion

Decommissioning offers many benefits to an organisation which can be both financially tangible and intangible. It is an essential part of managing ageing systems and a continuous process for ensuring an organisation reduces its costs, minimises complexity and can deliver to market in a timely manner. However, decommissioning is not always easy and does not come for free. It is not a once-off effort. It is not a quick and easy opportunity to reduce cost.

http://doi.org/10.18080/ajtde.v5n2.109

Rather, decommissioning is an ongoing evolutionary process of modernisation that should be an innate capability of any organisation. Indeed, organisations such as NSW State Records, (2015) understand the multidimensional benefits of decommissioning value include obsolescence in their design processes to ensure a system's end-of-life is planned and easy.

Lastly, efforts by management and staff to decommission ageing systems should not require an act of heroism. The act of decommissioning should be seen as delivering as much value as turning on new systems and as such should be recognised, rewarded and celebrated for the benefits it delivers.

Acknowledgements

This industry report has been written and reviewed with the help and grateful appreciation from the following people and organisations:

- Colin Fairweather – CIO, City of Melbourne
- Simon Weller – Enterprise Architect, City of Melbourne
- Steve Hodgkinson - CIO, Victorian Department of Health & Human Services
- Dr Kerry Hinton and Dr Leith Campbell – School of Engineering, University of Melbourne
- Matthew Lunn
- David Koetsier

References

DiCaprio, T. 2012, "Becoming Carbon Neutral – How Microsoft is striving to become leaner, greener, and more accountable", *Tamara DiCaprio - Microsoft Corporation*, June 2012. Available at http://www.microsoftgreen.com/2016/03/28/how-microsoft-is-empowering-a-modern-sustainable-workplace/

DiCaprio, T. 2015, "Making an Impact with Microsoft's Carbon Fee", *Tamara DiCaprio - Microsoft Corporation*, March 2015. Available at http://download.microsoft.com/download/0/A/B/0AB2FDD7-BDD9-4E23-AF6B-9417A8691CF5/Microsoft%20Carbon%20Fee%20Impact.pdf

Hormann, P. 2014, "Data Storage Energy Efficiency in the Zettabyte Era", *Peter Hormann and Dr Leith Campbell - Australian Journal of Telecommunications and the Digital Economy*, Vol 2, No 3 - October 2014. Available at http://telsoc.org/ajtde/2014-10-v2-n3/a51

Lambrechts, J. 2014, "Information Lifecycle Governance (ILG) - Maximise data value, reduce data growth, cost and risk", *Jan Lambrechts - Australian Journal of*

Telecommunications and the Digital Economy, Vol 2, No 3 - October 2014. Available at http://telsoc.org/ajtde/2014-10-v2-n3

Murphy, P. 2011, "Application Retirement – It's Time to put the Elephant in the Room on a Diet", *Phil Murphy with John Rymer and Sander Rose - Forrester Research*, February 2011. Available for purchase at https://www.forrester.com/report/Application+Retirement+Its+Time+To+Put+The+Elephant+In+The+Room+On+A+Diet/-/E-RES58077

NSW State Records, 2015, "Decommissioning Systems – The Benefits", *State Records Department, NSW Government*, December 18th, 2015. Available at https://futureproof.records.nsw.gov.au/decommissioning-systems-the-benefits/

Queensland CIO, 2016, "ICT Risk Matrix" Queensland Government Chief Information Office. Accessed on November 4th, 2016 https://www.qgcio.qld.gov.au/products/ict-risk-management/ict-risk-matrix

Wikipedia, 2016, "Asset Retirement Obligation" – *Wikipedia*. Accessed on October 8th, 2016 https://en.wikipedia.org/wiki/Asset_retirement_obligation

http://doi.org/10.18080/ajtde.v5n2.109