

An Analysis of Logical Network Distance on Observed Packet Counts for Network Telescope Data

Barry Irwin¹ and Richard J Barnett²
 Security and Networks Research Group
 Department of Computer Science
 Rhodes University
 Grahamstown, South Africa
 E-Mail: ¹b.irwin@ru.ac.za ²barnettrj@acm.org

Abstract—This paper investigates the relationship between the logical distance between two IP addresses on the Internet, and the number of packets captured by a network telescope listening on a network containing one of the addresses. The need for the computation of a manageable measure of quantification of this distance is presented, as an alternative to the raw difference that can be computed between two addresses using their Integer representations. A number of graphical analysis tools and techniques are presented to aid in this analysis.

Findings are presented based on a long baseline data set collected at Rhodes University over the last three years, using a dedicated Class C (256 IP address) sensor network, and comprising 19 million packets. Of this total, 27% by packet volume originate within the same natural class A network as the telescope, and as such can be seen to be logically close to the collector network.

The paper concludes with an evaluation of the hypothesis of there being a relation between the logical distance and volume of traffic received, and addresses some possible vulnerabilities in commonly used anonymization tools for such packet traces.

Index Terms—Internet Background Radiation, Network Telescope, Passive Monitoring

I. INTRODUCTION

INTERNET Protocol (IP) version 4 addresses can be seen as existing on a discrete continuum of integers from 0 to 2^{32} . It is only through common notation that they are most often represented in 'dotted-quad' decimal of the form 192.192.192.123. In their most basic form they can be considered to be a unsigned integer, and are processed as such by almost all network stacks, and exist as such in the native packet format [1], [2], [3], [4].

Consequently it is possible to calculate a difference between two IP addresses in order to determine their logical or numerical distance from each other within this continuum. This may be, and in most cases is, significantly different from their topological distance (the number of network hops between them) and any physical distance calculated between the locations of their geophysical manifestations.

The authors would like to acknowledge the financial support of Telkom SA, Business Connexion, Comverse SA, Stortech, Tellabs, Amatole, Mars Technologies and THRIP through the Telkom Centre of Excellence in the Department of Computer Science at Rhodes University.

The value of the computation of such a logical distance lies in its use in quantifying the network locality or nearness of two IP addresses. The ability to provide a quantification of the relative nearness of two addresses is something which is particularly useful in the analysis of network aware malware's scanning algorithms, and consequently the analysis of network traffic collected by IDS or Sensor networks such as that used in this study. In such networks it has been shown (as in Bykova and Osterman[5] and Moore et al. [6]) that a sensor network is likely to experience a notable bias towards traffic originating from networks that are numerically or logically close. This is largely due to the relatively primitive scanning and propagation algorithms employed by malware automata, and naive crackers. Most automata to date have evolved from the totally random scanning patterns of earlier incarnations and operate in two distinct modes near and far. The near scanning mode is employed first where networks numerically near to that of the infected host are scanned first, and then once this is complete a shift can be seen to a more randomized scanning of far address space. This is particularly well exhibited with worms such as Code Red and Code Red II , and also exemplified by Blaster as discussed by [7], [8], [9], [10].

A. Paper Organisation

The remainder of the paper is structured as follows, with Section II providing details on the data collection methodology, and equipment setup. Section III introduces the concept of a Distance Score - the calculated logical distance between two IP addresses. Section IV introduces a Hypothesis relating to the interpretation of this score. The analysis of the results of the calculation of this score on the sample data is discussed in Section V with a focus on IP address pairings with low distance scores. Some suggestions for future work that lead out of this research are discussed in section VI. Finally, Section VII presents conclusions to this work.

II. DATA COLLECTION

The data set used as the basis for this analysis has been collected by the authors as part of a long term project at

Algorithm 1 Dotted Quad to Integer Conversion

$$INT(A.B.C.D) = A * 256^3 + B * 256^2 + C * 256 + D$$

Rhodes University to collect and analyze background radiation on the Internet. Data consists of captured IP datagrams destined for a dedicated class C (/24) network, which logs the packets, and then discards them. As such there is no interaction at all, and it is generally difficult for an attacker to know if the packets have actually reached a destination or have been dropped elsewhere on the routing path. The time period analysed (August 2005 - April 2008) comprises 19 million packets. These datagrams have been loaded into a PostgreSQL RDBM for further analysis. While not all traffic captured by this network sensor can be classified as potentially malicious, the bulk of it can be shown to be so. Further details on Internet background radiation can be found in Pang *et al.*[11]. The remainder consists of backscatter traffic received from misconfigured hosts or as evidence of spoofing or denial of service attacks elsewhere on the Internet. One shortcoming of the current passive collection methodology is that only the first packet of a TCP connection is received. With the lack of the completion of the '3-way handshake' and the consequent flow of data, it is difficult in most cases to determine with certainty as to the potential purpose of the packet given the lack of payload data. UDP and ICMP packets do not suffer as badly from this restriction, but in the case of the former require extensive knowledge of the OSI layer 7 protocol in use in order to interpret further. The analysis performed in this paper is based on the entire data set, and has not attempted to discriminate based on protocol, or otherwise further classify the traffic as potentially malicious or backscatter. A second sub sample of the entire set comprising traffic originating from the 196.0.0.0/8 netblock is also used as this is the parent /8 network of the monitored /24 address block.

III. DISTANCE SCORE CALCULATION

The following section discusses the derivation of a formula to quantify the logical distance between two arbitrary IP addresses. The basic method for the conversion of a dotted quad IP address into an integer is to use the formula as shown in Algorithm 1. This is in effect what is used internally by POSIX system calls such as `atoi()` and inversely by `itoa()` in system libraries and networking stacks. It provides a baseline metric for preliminary assessment, from which further refinement can take place.

A. Evaluation

While the conversion discussed above gives is accurate, the actual distance between two addresses has little need of the least significant components since they are in most cases on the same network, or at least with in the same organisation or logical netblock. As such the conversion can be refined to being $A * 2^{24} + B * 2^{16} + C * 2^8$. (effectively omitting the last octet which usually specifies the host from the calculation shown in Algorithm 1). At this point, it is worth noting that while direct assignments to organizations of smaller than

/24 do exist they are relatively few and are from a legacy period prior to the establishment of the regional registries such as LACNIC and AFRINIC in 2002 [12], [13], [14]. Such assignments to organizations from their service provider are however much more common with assignments of /28 and /29 being prevalent for broadband connectivity. [15], [12], [16], [17]. The reduced form of the conversion proposed above logically clumps such networks as being part of the same higher level assigned network block, which is likely to be a direct assignment from a Regional Registry. Examples of using the baseline and the reduced formula are shown below:

Example 1: Addresses within the same natural subnet (255.255.255.0)

IP Address A : 192.168.149.67 (3232273731)

IP Address B : 192.168.149.254 (3232273918)

This gives a natural difference of 187

Using the second method provided above the difference can be shown to be 0 when the least significant octet is omitted.

Example 2: Differing network addresses

IP Address A: 146.231.123.15

IP Address B: 209.67.212.202

Here the Difference can be shown to be 1 046 239 675 or approximately 1 billion addresses apart (1.046×10^9). Removing the least significant byte from the calculation gives 1 046 239 488 which is only 187 ($210 - 15$) different from the answer obtained by the method shown in Example 1. This difference is in itself is insignificant as it accounts for less than $\frac{1}{1000}$ th of a percent of the difference calculated. For the purposes of the remainder of this work the full four octets were utilised in the calculation, as the omission can be shown to have negligible benefit, and can actually prove detrimental when taken into consideration with the re-factoring discussed in the following section. When one takes into account the large values which can be obtained as the result of natural subtraction of the components (using either algorithm) the need to be able to reduce this to a more comprehensible number is of some interest. The method used to achieve this is discussed in the following section.

B. Reduction of the raw difference score

The ideal was to be able to reduce the potentially large range of differences from $\pm(0 \rightarrow 2^{32} - 1)$ to a more comprehensible range. Taking the Log_{256} of the absolute integer difference was decided on as this provided a score in the somewhat reduced and more comprehensible range of $0.0 \leq 4.0$. This was chosen as it represents a concise, tangible range along with relating numerically well to the manner in which IP addresses and their integer representations are related. The resultant values can be interpreted as follows:

$0 \leq 1$ Addresses lie on the same network (there are 255 or less individual addresses between the two)

$1 \leq 2$ Addresses lie within two /16 networks (65535 IP addresses) of each other. This may not necessarily be a block of addresses lying on a contiguous natural (2^n) boundary.

$2 \leq 4$ Addresses lie elsewhere on the Internet, with the values approaching 4.0 as the distance increases.

Table I
COMMON $IP\Delta$ VALUES

$IP\Delta(\log_{256})$	Raw Score	# of /24 Nets	# of /16 Nets
1	1	1	0
1.38	2048	8	0.03
1.5	4096	16	0.06
1.88	32768	128	0.5
2	65536	256	1
2.13	131072	512	2
2.25	262144	1024	4
2.38	524288	2048	8
2.5	1048576	4096	16
2.63	2097152	8192	32
2.75	4194304	16384	64
2.88	8388608	32768	128
3	16777216	65535	256

A practical maximum lies closer to 3.9759 when one accounts for the maximal distance between 0.0.0.0 and 224.0.0/4 which is the start of multicast address space [18]. When interpreting these values, it is important to note that they are unsigned, and as such there is no directionality associated with the distance, while this has its possible disadvantages, it obviates the problem of the ordering of the IP addresses in the calculation. Thus when using the score for measurement it in effect is a score of $\pm\Delta$, providing a range on either side of the target address. The resultant formula for the calculation of the Distance Score ($IP\Delta$) between two IP addresses can be defined as in Algorithm 2. This algorithm could be extended by storing the sign bit of the initial calculation and then re-applying it post calculation of $IP\Delta$. Table I shows some common boundary values, which can be used for providing a finer grained interpretation of the results of the calculations.

Algorithm 2 Calculation of IP Distance Score: $IP\Delta$
 $IP\Delta = \text{Log}_{256}(\text{ABS}(\text{INT}(IP_A) - \text{INT}(IP_B)))$

IV. HYPOTHESIS

The authors hypothesise that due to the relatively crude scanning and propagation algorithms [18], [7], [19], [20], [21] implemented by much of the malware automata seen on the Internet, there will be a natural bias of the telescope to see proportionally higher traffic from numerically closer networks. As such the Distance Score is calculated for two large samples from the Rhodes Telescope Data Set (as previously described) and a number of graphical and statistical analyses performed. It is still to be ascertained if this hypothesis is a side effect of smaller telescope sizing such as the /24 used in the case of the authors sensor. The rationale for this is that when looking at larger telescope data sets such as those provided by the CAIDA Backscatter-2007 Dataset [22] the size of their telescope is reported to be a /8 network equating to 1.67 million (2^{24}) target addresses. If set up in a true telescope fashion, no radiation should be emitted from the collector network. What this does, is in effect removes the hypothesized bias since its neighboring networks are likely to be more highly dispersed than in the case of a /24 sensor.

A. Data Sets

The analysis to validate the hypothesis was initially performed on a data set of 68 217 distinct addresses located within the same /8 netblock as the network telescope. This represented the monitored IP addresses closest to the telescope, from a naïve numerical perspective, and all under the control of the AFRINIC Regional Registry (RR) [23]. As such many of these are ostensibly located somewhere on the African continent and surrounding islands, further suggesting geographical closeness. It is worth noting that as of the time of writing (April 2009), there is still very poor inter-connectivity between most African countries, with topological paths in most cases having to traverse via European or US based peering points [24]. This promises to change in the future given the installation of several submarine cable projects. Given this poor degree of close inter-connectivity, the numerical and geographical closeness of these addresses cannot be seen to translate to topological closeness on the Internet. However it must be further noted that a significant portion of this address space is contained and operated within South Africa. South African IP space does have good inter-connectivity between providers through public exchanges facilitated by the Internet Service Providers Association of South Africa (ISPA) [25], and private peering between Tier-1 providers, and thus are generally topologically close. This dataset comprises 27% of the total by packet count of the full dataset during the three year observation period, whereas only occupying $1/256^{\text{th}}(\pm 4.0\%)$ of the total Internet IPv4 address space, showing a relative high PACKET:ADDRESS density ratio. A second dataset comprising the same data for all 1 237 739 distinct addresses monitored and including the first dataset was also used. For the purposes of the analysis of the two datasets above, the distance score was computed from the source address of the captured packet to the midpoint of the telescope network a.b.c.128 (mapping this to the IP_B value in the calculation), using the technique described in Algorithm 2. This was felt to be a good midpoint balance as it represented the midpoint of the observed network, which had a fairly even distribution of traffic.

Datasets were extracted from the archive database of the traffic data with the appropriate calculations for packet count, and scoring being implemented within the database. PostgreSQL provides for the native mathematical manipulation of the `inet` and `cidr` data types (and the consequent resulting integer values) which simplified the processing required.

V. ANALYSIS

This section discusses the interpretation of graph data plotted using the packet count and the distance scores calculated from the collected datasets. A radar plot of the Log_{10} packet count vs the $IP\Delta$ score values is shown in Figure 1. This is significant in that it shows the apparent correlation at the lower end of the distance score to high packet counts. This image contains the 250 closest recorded addresses to the network telescope as determined by the distance score ($IP\Delta$). The Log_{10} of the packet count has been taken in order to have the two series on relatively comparable axes.

Figure 1. Radar plot of distance vector score ($IP\Delta$) vs. \log_{10} Packet Count

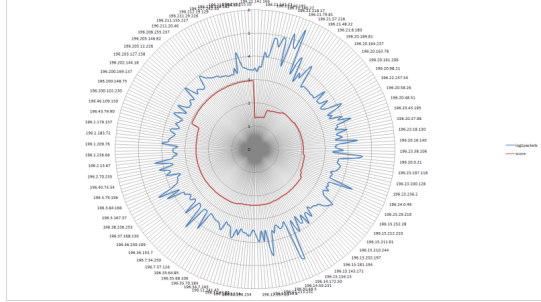


Figure 2. Zoomed plot showing first 38 closest IP nodes by $IP\Delta$ order.

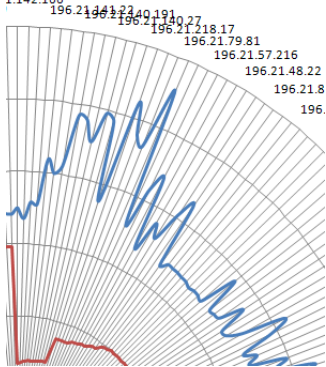


Figure 2 shows an annotated closeup of the lead portion of the plot containing the first 38 address nodes (those with a distance score of <2.00), and as such very close, representing a minimal distance of 1 753 addresses away, to a maximum of 62 692. This effectively means that at a minimum the closest nodes were within 6.8 natural /24 netblocks of the telescope midpoint, with a maximum of 244 /24 netblocks distance (or in effect approximately one /16 netblock away). For this dataset, the mean distance was 3.67469 with a standard deviation of 0.2980969. This data is further illustrated in Figure 3 where the general trend for packet count by source address can be seen to decrease with the corresponding increase in $IP\Delta$ value. The ratio of $IP\Delta : \log_{10}PacketCount$ is also shown. This image shows the same dataset as Figure 1. Further to this Figure 4 shows the percentage contribution to the total packet count of the closest 250 nodes, along with the $IP\Delta$ Score.

Examining a Hilbert curve plot of the entire dataset, Figure 5 shows a trimmed portion of the resultant image, showing the section of 196.0.0.0/8 (administered by AFRINIC) [23] and the upper portion of 195.0.0.0/8, administered by RIPE. 197.0.0.0/8 has been omitted as it was unallocated [23] during the period in question, and no traffic has been recorded claiming to originate from this region. The heat-map colouring shows the number

Figure 3. Linear plot of Distance Scores vs Packet count and computed ratio

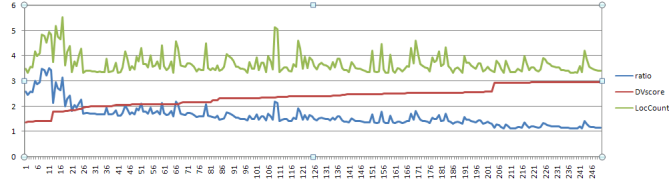


Figure 4. Percentage Contribution to packet count of top250 nodes

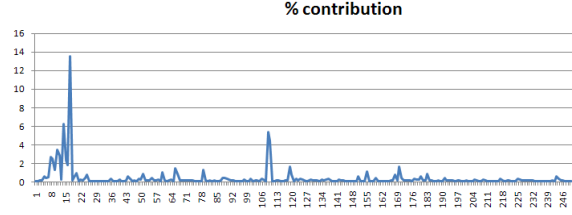


Figure 5. Heat-map showing 196.0.0.0/8 whose assignment is controlled by AFRINIC.



The Sensor Network is a class C network that lies within this range, and this this represents the closest IP addresses.

of packets received, with red showing the highest counts. In this image each pixel represents a discrete /24 netblock. Most of 196.128.0.0/9 can be seen to be unallocated (and this can be verified from BGP routing tables). The upper quartile of the 195.0.0.0/8 block (i.e. 195.192.0.0/10) can be seen to show as significantly warmer than the other shown elements within this range. It is proposed that this is due to the closeness of this range to that of the telescope sensor network, with a maximal distance of $\Delta 2.8$, or approximately 7 million IP addresses of the telescope. Similar results would be expected in the lower quartile of 197.0.0.0/8 if it were in use. A full view of the captured data can be seen in Figure 6, where the colouring of the 196.0.0.0/8 block should be noted as having a higher packet count density. The only other significant areas are address blocks belonging to APNIC (218.0.0.0/7 and 60.0.0.0/7) although these are still significantly less dense than the 196.0.0.0/8 block. Also worth noting is the high concentration in a small area of 41.0.0.0/8 which is also under the control of AFRINIC. This small concentration in the central region of this block can be identified as the DSL IP ranges used by Telkom Internet and its resellers. This last point shows that geographically close networks (provided there is good interconnect) may also have higher than average packet counts.

Figures 7a and b respectively show the networks within 196.0.0.0/8 being advertised via BGP (obtained from a Tier-

by most malware to date. The observation of over a long temporal baseline of this phenomenon, lends some credence and validity to the conventional wisdom. Of somewhat larger value is that understanding the relationship between the two variables under study in this paper will lead to an overall better understanding of the complexities in modeling Internet Traffic, and the analysis of the backscatter commonly seen by any connected host on the Internet.

A more sinister ramification of this is that given suitably sized datasets, researchers may be able to reverse the current blinding mechanisms used for anonymization of packet traces for public release such as those made available by CAIDA or other organisation. The ability for one to determine the location of a masked network telescope through such statistical and logical inference may well allow for the future pollution of data or injection of chosen payloads in to the datasets with the intent to exploit vulnerabilities in common packet processing tools such as tcpdump [1] and Wireshark [29]. tcpdump and Ethereal (Wireshark's predecessor) have had some well publicized vulnerabilities which could be exploited, especially since they are tools usually run with administrative privileges. This potential for attack is however only feasible for certain data types, most of which are not commonly seen in network telescope packet captures.

REFERENCES

- [1] "tcpdump," [online] <http://www.tcpdump.org/>. [Online]. Available: <http://www.tcpdump.org/>
- [2] J. Postel, "Internet Protocol," RFC 791 (Standard), Sep. 1981, updated by RFC 1349. [Online]. Available: <http://www.ietf.org/rfc/rfc791.txt>
- [3] W. R. Stevens, *TCP/IP illustrated (vol. 1): the protocols*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1993.
- [4] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," RFC 2474 (Proposed Standard), Dec. 1998, updated by RFCs 3168, 3260. [Online]. Available: <http://www.ietf.org/rfc/rfc2474.txt>
- [5] M. Bykova and S. Ostermann, "Statistical analysis of malformed packets and their origins in the modern internet," in *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement*. New York, NY, USA: ACM, 2002, pp. 83–88.
- [6] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the slammer worm," *IEEE Security and Privacy*, vol. 1, no. 4, pp. 33–39, 2003. [Online]. Available: <http://portal.acm.org/citation.cfm?id=939830.939954>
- [7] Z. Chen, C. Chen, and C. Ji, "Understanding localized-scanning worms," in *27th IEEE International Performance Computing and Communications Conference, IPCCC 07*, New Orleans, LA, 2007, pp. 186–193.
- [8] Z. Chen and C. Ji, "Optimal worm-scanning method using vulnerable-host distributions," *International Journal of Security and Networks: Special Issue on Computer and Network Security*, vol. 2, no. 1/2, 2007.
- [9] C. C. Zou, W. Gong, and D. Towsley, "Code red worm propagation modeling and analysis," in *CCS '02: Proceedings of the 9th ACM conference on Computer and communications security*. New York, NY, USA: ACM Press, 2002, pp. 138–147.
- [10] S. Qing and W. Wen, "A survey and trends on internet worms," *Computers & Security*, vol. 24, no. 4, pp. 334–346, Jun. 2005. [Online]. Available: <http://www.sciencedirect.com/science/article/B6V8G-4G1GF4C-1/2/082f72527f1fddd36992b2a8e6fabfdf>
- [11] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of internet background radiation," in *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2004, pp. 27–40.
- [12] E. Gerich, "Guidelines for Management of IP Address Space," RFC 1466 (Informational), May 1993, obsoleted by RFC 2050. [Online]. Available: <http://www.ietf.org/rfc/rfc1466.txt>
- [13] K. Hubbard, M. Koster, D. Conrad, D. Karrenberg, and J. Postel, "Internet Registry IP Allocation Guidelines," RFC 2050 (Best Current Practice), Nov. 1996. [Online]. Available: <http://www.ietf.org/rfc/rfc2050.txt>
- [14] V. Fuller and T. Li, "Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan," RFC 4632 (Best Current Practice), Aug. 2006. [Online]. Available: <http://www.ietf.org/rfc/rfc4632.txt>
- [15] P. Tsuchiya, "On the assignment of subnet numbers," RFC 1219 (Informational), Apr. 1991. [Online]. Available: <http://www.ietf.org/rfc/rfc1219.txt>
- [16] IANA, "Special-Use IPv4 Addresses," RFC 3330 (Informational), Sep. 2002. [Online]. Available: <http://www.ietf.org/rfc/rfc3330.txt>
- [17] Z. Albanna, K. Almeroth, D. Meyer, and M. Schipper, "IANA Guidelines for IPv4 Multicast Address Assignments," RFC 3171 (Best Current Practice), Aug. 2001. [Online]. Available: <http://www.ietf.org/rfc/rfc3171.txt>
- [18] J. Nazario, *Defense and Detection Strategies against Internet Worms*. Norwood, MA, USA: Artech House, Inc., 2003.
- [19] Z. Chen and C. Ji, "Measuring network-aware worm spreading ability," in *26th Annual IEEE Conference on Computer Communications IEEE INFOCOM 2007*. Anchorage, Alaska, USA: IEEE, 6–12 May 2007.
- [20] C. C. Zou, D. Towsley, and W. Gong, "On the performance of internet worm scanning strategies," *Performance Evaluation*, vol. 63, no. 7, pp. 700–723, Jul. 2006. [Online]. Available: <http://www.sciencedirect.com/science/article/B6V13-4H2FXPJ-1/2/70426d9f159f2b4822a6a7ce3989f75d>
- [21] M. Liljenstam, D. M. Nicol, V. H. Berk, and R. S. Gray, "Simulating realistic network worm traffic for worm warning system design and testing," in *WORM '03: Proceedings of the 2003 ACM workshop on Rapid malware*. New York, NY, USA: ACM Press, 2003, pp. 24–33.
- [22] C. Shannon, D. Moore, and E. Aben, "The CAIDA Backscatter-2007 Dataset - January 2007 - November 2007, (collection)," Online, CAIDA Network Telescope Project - Backscatter, 2007, support for the Backscatter-2007 Dataset and the UCSD Network Telescope are provided by Cisco Systems, Limelight Networks, the US Department of Homeland Security, the National Science Foundation, DARPA, Digital Envoy, and CAIDA Members.
- [23] Internet Assigned Numbers Authority (IANA), "Ipv4 global unicast address assignments," [online] <http://www.iana.org/assignments/ipv4-address-space>, 2008-05-27. [Online]. Available: <http://www.iana.org/assignments/ipv4-address-space>
- [24] African Internet Exchange Point Task Force. [Online]. Available: <http://afix.afrispa.org/>
- [25] Internet Service Providers Association of South Africa, "Johannesburg internet exchange (jinx)," [online] <http://www.ispa.org.za/jinx/index.shtml>. [Online]. Available: <http://www.ispa.org.za/jinx/index.shtml>
- [26] B. Irwin and N. Pilkington, "High level internet scale traffic visualization using hilbert curve mapping," in *VizSEC 2007 Proceedings of the Workshop on Visualization for Computer Security*, ser. Mathematics and Visualization, G. Conti, J. R. Goodall, and K.-L. Ma, Eds. Springer Berlin Heidelberg, May 2008, pp. 147–158.
- [27] The Measurement Factory, "Ipv4 heatmap software," [online] <http://maps.measurement-factory.com/software/>, 2008.
- [28] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2007, ISBN 3-900051-07-0. [Online]. Available: <http://www.R-project.org>
- [29] "Wireshark," [online] <http://wireshark.org/>. [Online]. Available: <http://www.wireshark.org/>

Mr Barry Irwin is currently completing his PhD research on the use of Passive Sensors as a means for inferring hostile network activity on the Internet. He holds a MSc in computer Science from Rhodes University, and has research interests in Information Security to modern IP networks particularly adaptive automated network defence and early warning systems, and data visualisation techniques to support these.