Measuring network change: Rényi cross entropy and the second order degree distribution

Edward F. Harrington

Intelligence, Surveillance and Reconnaissance Division,
Defence Science and Technology Organisation,
Lock Bag 5076, Kingston,
Canberra, ACT 2604, Australia
edward.harrington@dsto.defence.gov.au

Abstract. Being able to detect temporal changes of a network's topology can prove useful for the effective management of networks. Recent work on network models has highlighted potential limitations of the degree distribution as a unique representation of a network. If the representation of the network is likely not to be unique then it makes it difficult to use for change detection. Motivating the consideration of an alternative approach which may improve the uniqueness of a network's representation is the second order degree distribution, defined as the distribution of the degree product of the edge-paired vertices of the network. Intuitively, this distribution captures the nature of the connected vertices within a network. When comparing two observations of a network, at two different points in time, their estimated distributions can be used to see if there is any change in the nature of the network's connections. To get an empirical measure of that temporal change the cross entropy of the two estimated distributions was used. Experiments were conducted to study the ability of this simple approach in detecting network variability, and therefore whether or not a network has changed.

1 Introduction

Advancements in communications technology and their networks, particularly the Internet, has led to a lot of research into trying to understand network growth behaviors. We consider the approach of having several snapshots of a network at different sequential points in time. Enabling us to gain an understanding of how network evolves as a function of time. Each network observation is modeled as an undirected graph $G(V, E)$ with N vertices (nodes) V, and M edges E (links). So, a network can be viewed as a time series of graphs $G_1, G_2, \ldots, G_t$. We are interested in measuring network change using this time series, specifically the temporal change in the topology of a network. Detecting temporal change of a network enables better understanding of a network's dynamics, thus enabling the better design of future networks and their protocols.

Several authors [4, 5, 9] have observed that the degree distribution of the topology of the *autonomous system* (AS) that constitutes the Internet can be

approximated using a power-law model. The degree is defined as the number of edges connected to a particular vertex of the network's graph. The AS degree distribution can be used to represent the Internet, as the Internet consists of tens of thousands of loosely connected AS networks. As noted by [8], one limitation of using the degree distribution is that it is not necessarily a unique representation of a network. This is evident by the fact that there exist methods for re-wiring networks with the same degree distribution. As it is improbable that every network will have a unique degree distribution to represent it, we propose in this paper the use of an alternative distribution which is more likely to be unique compared to the degree distribution. Proposed is the use of a second order degree distribution to represent each network in time. This distribution is the degree product of the edge-paired vertices of the network. The proposed representation has the advantage that it captures the nature of the attachment, or connectivity, of the network. In [10] it was shown that the correlation of the network degrees can determine the attachment of a network. The nature of the attachment is determined by whether or not the network has *assortative mixing* or *disassortative mixing*. The assortative mixing is defined by the linkage of high-degree vertices to other high-degree vertices, and the disassortative mixing is defined by the linkage of high-degree vertices to low-degree vertices. Like the correlation, the second order degree distribution gives some indication of assortative mixing because a heavy tail in its distribution is indicative of assortative mixing.

After determining the second order degree distribution of each observation of a network we require a measure of comparison between observations. Recently, in [6] they proposed a measure of change for streamed data using the Kullback-Leibler (K-L) distance [1]. The streamed data consisted of a set of features like TCP connection time and total down-load time. In their method a sliding window is applied to the data at each node of the network and the K-L distance between two successive windows is then determined. Also, bootstrapping is applied to the windowed data (random re-sampling is done) to get a statistical measure of significance. Their approach measures the change at each vertex of the network. The difference between the method in [6] and the method proposed in this paper is that the method of this paper measures the change of the network topology as a whole, and not just one vertex at a time. Also, the measure of change we consider is symmetrical; unlike the K-L distance which is sensitive to the comparison's order (see Section 3).

The rest of the paper is organized as follows. Section 2 provides details of the representation of each observation of the network, including details of degree and second order degree distributions. Section 3 presents the proposed method for detecting topological change. Sections 4 and 5 discuss the results of experiments conducted on simulated and AS-topology data. In Section 6 we conclude with a summary of results, and present ideas for further research.

[1] Strictly speaking it is not a distance, as it does not obey the triangular inequality.

2 Representation

We consider a sequence of observations of a network G, where the observations are not necessarily sampled at fixed intervals of time, resulting in the series $G_1, G_2, \ldots, G_t$. For each observation G_j has corresponding degrees $\mathbf{w}(G_j) = \{w_1, w_2, \ldots, w_N\}$, where w_i is the number of degrees for vertex i and N being the number of vertices. From $\mathbf{w}(G_j)$ the degree distribution $\mathbf{p}_{w_i}(G_j)$ is estimated using a histogram denoted by $\tilde{\mathbf{p}}_{w_i}(G_j)$. The degree distribution provides one way of representing a graph as a probability distribution, and subsequently the ability to provide a measure of the topology of the network.

An alternative to the degree distribution is to use the distribution of the probabilities of a network's edges (links) rather than a network's degrees. An edge of a network, denoted by $e_{i,j}$, being the link between the network vertices i and j. If the two vertices i and j are connected then $e_{i,j} = 1$ else $e_{i,j} = 0$. We do not consider the (estimation of the) probability of edges directly here in this paper, that is $\mathbf{p}_{e_{i,j}}$, though the measures discussed in the rest of this paper are applicable to probability of edge approaches as well. One reason for using the degree distribution compared to the edge probabilities is that it is far easier to compare two different observations in time of a network using the degree distribution. The reason it is easier is because it avoids dealing with the possibility that the observations of a given network may have non matching vertices, due to a birth/death process of vertices. For example one network may have the vertices 1,4,5,8 and another has 1,5,8,15. Clearly this makes it hard to compare these two networks.

An associated reason for using the degree distribution is that it avoids the problem of estimating the probability of edges when the network may be non-stationary making sampling an issue. The pragmatic issue of sampling is that to estimate the probability of a given edge we must count the occurrence of that edge over some window of time in which we can safely assume that the network is stationary; this seems hard or even impossible, given the goal itself is to determine whether the network is stationary.

While the degree distribution avoids the sampling issue of the edge probability approach, as discussed in [8], the degree distribution has the limitation that it is possibly not a unique representation of a network. The recent work of Newman [10] indicated that to model a network accurately some information regarding the nature of attachment is important, as well as making sure the model's degree distribution observes a power law [1, 11]. To address the limitation of the degree distribution, and to capture in the representation the type of attachment the second order degree distribution is used. The second order degree is defined by $\mathbf{w}_x\mathbf{w}_y(G) = \{w_i w_j : (i,j) \in E(G)\}$ where x, y are vertices of network G, and $E(G)$ is the set of all edges of network G. So, the second order degree distribution is $\mathbf{p}_{w_i w_j}(G)$, and its estimate is denoted by $\tilde{\mathbf{p}}_{w_i w_j}(G)$. If most of the second order degrees of $\mathbf{w}_x\mathbf{w}_y(G)$ are large then this is reflected by a heavy tailed second order degree distribution, and this is indicative of the presence of assortative mixing within the network.

3 Measuring change

In this section we define and describe the properties of Rényi entropy and cross entropy measures. Also, we described the use of cross entropy for measuring a network's topological change.

The Rényi entropy [12] measure of a proper probability distribution[2] of order α is

$$H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha}\log_2 \sum_k p_k^\alpha, \tag{1}$$

where $0 < \alpha < 1$. The Shannon entropy measure is a special case of the Rényi entropy for $\alpha \to 1$. From (1) the (Rényi) cross entropy of order α of is derived

$$I_\alpha(\mathbf{p}, \mathbf{q}) = \frac{1}{1-\alpha}\log_2 \sum_k \frac{p_k^\alpha}{q_k^{\alpha-1}}, \tag{2}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two discrete distributions. The K-L (Kullback-Leibler [7]) distance is a special case of the cross entropy of (2) for when $\alpha \to 1$. One important property of the cross entropy is that if $\mathbf{p} = \mathbf{q}$ then $I_\alpha = 0$. As mentioned in the introduction, we are particularly interested in a measure which is symmetric, i.e. $I_\alpha(\mathbf{p}, \mathbf{q}) = I_\alpha(\mathbf{q}, \mathbf{p})$. If we chose $\alpha = 0.5$ in (2) then the cross entropy is symmetric. Throughout the rest of this paper when referring to the cross entropy we mean the symmetric case of (2)

$$I_{0.5}(\mathbf{p}, \mathbf{q}) = 2\log_2 \sum_k \sqrt{p_k q_k}. \tag{3}$$

To detect any change in a network we compare successive observations from the series of observations of the network. To determine if there is a change at time t we first compare the second order degree distribution $\tilde{\mathbf{p}}_{w_i w_j}(G_t)$ with the estimated distribution of the previous network's observation $\tilde{\mathbf{p}}_{w_i w_j}(G_{t-1})$ by using (3), i.e.

$$I_{0.5}(\tilde{\mathbf{p}}_{w_i w_j}(G_t), \tilde{\mathbf{p}}_{w_i w_j}(G_{t-1})) = 2\log_2 \sum_k \sqrt{\tilde{\mathbf{p}}_{w_i w_j}(G_t)\tilde{\mathbf{p}}_{w_i w_j}(G_{t-1})}. \tag{4}$$

To test whether there is a change between observations we use

$$I_{0.5}(\tilde{\mathbf{p}}_{w_i w_j}(G_t), \tilde{\mathbf{p}}_{w_i w_j}(G_{t-1})) \begin{array}{c} \text{Change} \\ < \\ > \\ \text{No Change} \end{array} \eta. \tag{5}$$

The choice of the threshold η is network dependent, so there is no one setting of η that we can prescribe without empirical evidence. The reliability of this change detection is also dependent of the accuracy of the estimate $\tilde{\mathbf{p}}_{w_i w_j}(G_t)$. That is intuitively the larger the network the more accurate we expect the estimate will be. Therefore, this method is more suitable for large networks (vertices in their thousands rather than hundreds).

[2] If $\mathbf{p}$ is a proper probability distribution then p_i the probability of event i possible distinct events has $\sum_i p_i = 1$, where $i = 1, \ldots, N$.

4 Simulated network

We start the experiments with a simulated network based on the *general model of random graphs* (GRG) [2]. The GRG is formed from the expected degrees of the graph vertices, $\mathrm{E}(w_i)$ for $i = 1, \ldots, N$. The edges are formed between vertices i and j at random and in proportion to the product $\mathrm{E}(w_i)\mathrm{E}(w_j)$. We used a power-law degree distribution $p_k = 1/k^\beta / \sum_{j=1}^{N} 1/k^\beta$ and randomly allocated a degree to each vertex such that the average of the vertex degree across the whole network is $N \times p_k$ which is equivalent to $\mathrm{E}(w_i)$, where $\beta = 2.1$ and degree k.

The reason for using a simulated graph, like GRG, is that it is possible to control the amount of difference between observations, and therefore the corresponding cross entropies. To create Figure 1 (a) we simulated change by generating a GRG for graph G_{t-1}, and formed G_t from G_{t-1} by keeping the vertex degrees less than a cut-off value. We therefore created differences in the tails of the successive distributions. To estimate distributions a histogram with a fixed number of bins (2000) was used, as this matched the number of vertices which was set at N=2000. Also, a histogram with this number of bins was a fair compromise between speed and approximation error. Figure 1 (b) was formed in essentially the same way as Figure 1 (a) only that rather than using a cut-off on the degree w_i to form G_t the cut-off is applied to the degree product $w_i w_j$.

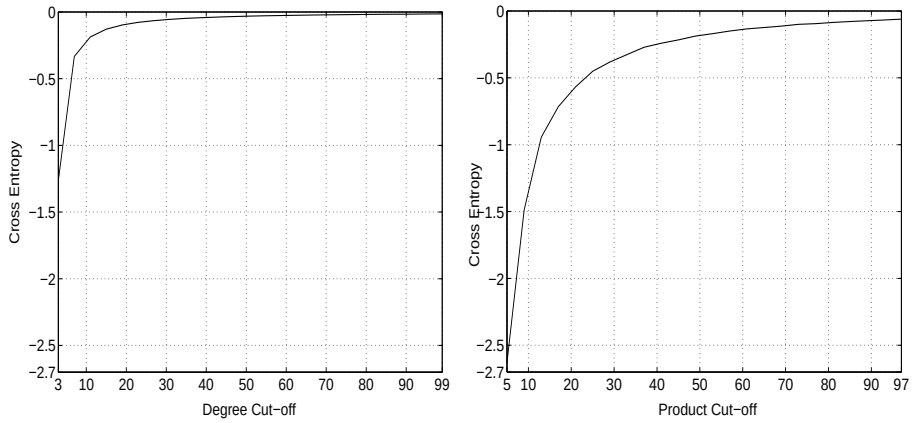
By comparing (a) and (b) of Figure 1 it is evident that there is a larger variation in the cross entropy values over the tested cut-off values of the second order degree distribution. Intuitively it makes sense that the second order degree distribution will have more spread in value compared to the degree distribution, because it is a product of degrees. This difference in cross entropy values indicates in this simulation that the second degree distribution has a larger range of change than the degree distribution.

5 AS-topology

The data set used is the well known AS-graphs, the Autonomous System topology of the Internet as collected by the University of Oregon Route Views Project [3]. AS-graphs was first studied in [4] where it was shown that AS-graphs exhibits a highly variable degree distribution. By highly variable they meant that the degree distribution is heavy tailed. A number of authors [4, 9] have suggested that the tail of the degree distribution can be crudely modeled by a power-law approximation. Since this is a data set with high variability it makes it an interesting data set to study with a second order degree distribution because it is a natural measure of variability, especially with regard to attachment (connectivity).

The AS-graphs (Autonomous Systems topology of the Internet) data consists of a time series of 14 observations (recordings) ranging in roughly 3 monthly intervals from 1997/11/08 to 2001/03/16. The edges of the observed graph were formed from the physical links between two Autonomous Systems, where each

[3] Available at `http://www.cosin.org/extra/data/internet/nlanr.html`.



(a) Degree distribution. (b) Second order degree distribution.

Fig. 1. Cross entropy between the whole distribution and the part of the same distribution less than the degree and product cut-offs.

vertex of a graph was an Autonomous System. Each observed graph (time series observation) was an undirected graph with all self-loops and parallel edges removed.

We calculated the cross entropies of neighbouring distribution estimates using (4), where histograms were used to form the distribution estimates. For determining the degree distribution the degree of each vertex in the graph was determined using a histogram, and then the final distribution was produced by passing the result to a second histogram where the number of bins was set equal to the maximum degree of all the vertices of the graph. In the case of the second order degree distributions a similar process was followed: firstly, the degree for each vertex in the graph was determined using a histogram; secondly, a histogram was created using the degree products of all the links of the graph $\mathbf{w}_x\mathbf{w}_y(G_t) = \{w_i w_j : (i,j) \in E(G_t)\}$ to produce $\tilde{\mathbf{p}}_{w_i w_j}(G_t)$, where the number of bins of the second histogram was set to the maximum second order degree $= \max(\mathbf{w}_x\mathbf{w}_y(G_t))$.

Figure 2 shows the following measurements performed on the AS-graphs time series: cross entropy of (4); average degree difference $\tilde{w}(G_t) - \tilde{w}(G_{t-1})$, where the average degree is given by $\tilde{w}(G) = \frac{\sum_{i \in V(G)} w_i}{|V(G)|}$ with $V(G)$ being the set of all vertices of graph G, and $|V(G)|$ as the cardinality of $V(G)$; average second order degree difference $\widetilde{w_i w_j}(G_t) - \widetilde{w_i w_j}(G_{t-1})$, where the average second order degree is given by $\widetilde{w_i w_j}(G) = \frac{\sum_{(i,j) \in E(G)} w_i w_j}{|E(G)|}$; degree variance difference $\hat{w}(G_t) - \hat{w}(G_{t-1})$, where the degree variance is given by $\hat{w}(G) = \frac{\sum_{i \in V(G)} (w_i)^2}{|V(G)|} - \tilde{w}(G)^2$;

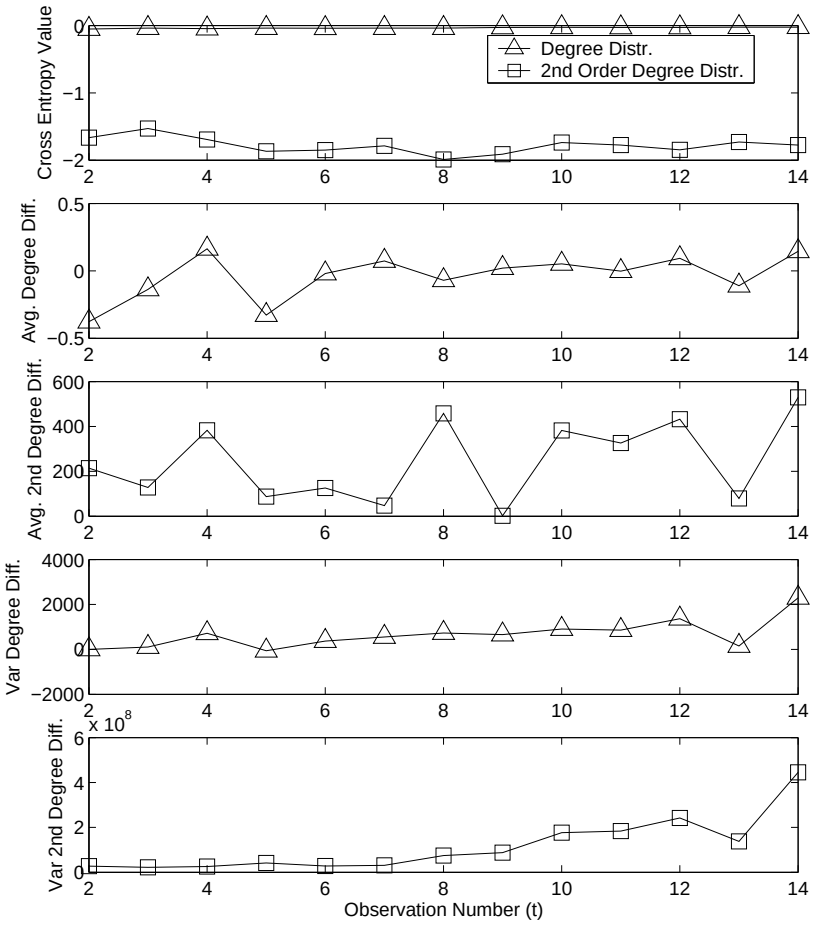


Fig. 2. Measurements of the AS-graphs time series (3 monthly intervals) at http://www.cosin.org/extra/data/internet/nlanr.html.

second order degree variance difference $\widehat{w_i w_j}(G_t) - \widehat{w_i w_j}(G_{t-1})$, where the second order degree variance is given by $\widehat{w_i w_j}(G) = \frac{\sum_{(i,j)\in E(G)} (w_i w_j)^2}{|E(G)|} - \widehat{w_i w_j}(G)^2$.

There are a number of observations to be made from results of Figure 2. The cross entropy values of the second order degree have significantly larger changes in value compared to the degree distribution. Noting that when the distributions are identical the cross entropy will be zero; the greater the difference in the distributions the more negative the resulting cross entropy will be. This difference between second order and first order degree histograms is support by the average and variance measurements. We notice for the second order degree results that the largest change in the time series was from time period $t-1=7$ to $t=8$ (observation 8 of Figure 2). We see from Figure 2 that this change in

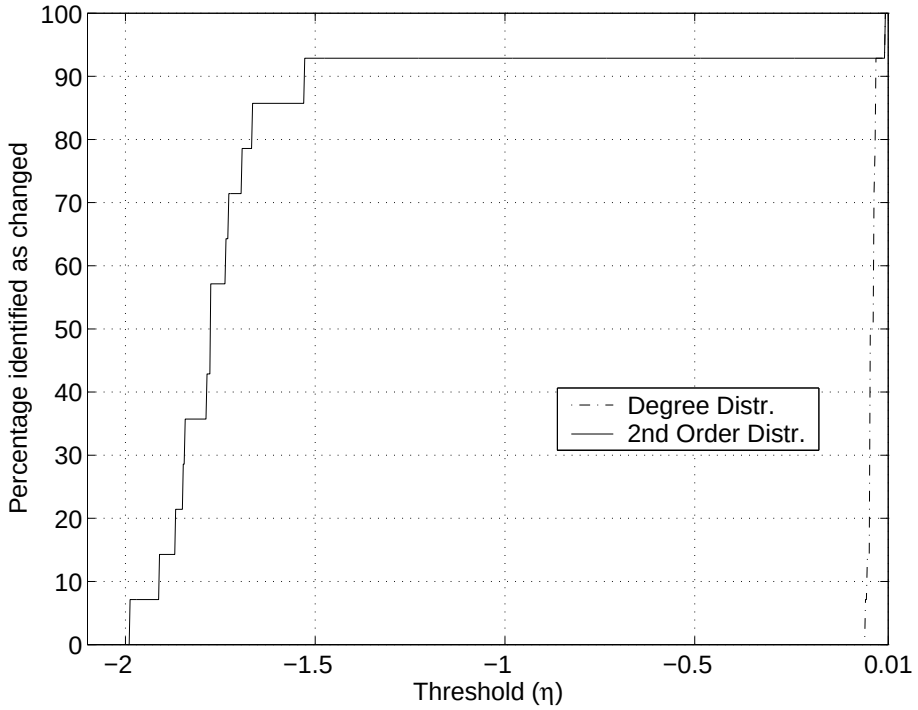


Fig. 3. Results of applying the change test of (5) to AS-graphs for various settings of the threshold η.

cross entropy is correlated with a sudden jump in both the average and variance of the second order degree distributions, which is indicative of an increase in the assortative mixing of the network.

Figure 3, shows the percentage of observations of the degree and second order degree distributions which were detected as having changed for various thresholds, η of (5). We see in Figure 3 that the spread of cross entropies is larger for the second order degree distribution compared to the first order degree distribution. This indicates that in the case of AS-graphs, and possibly other "large" networks, it is easier to identify changes in topology using the second order degree distribution.

5.1 Daily interval

To study the daily changes of the AS-level topology we used a modified version of the link data located at `http://irl.cs.ucla.edu/topology/data/2004.01`, the daily AS-level links for the month of January 2004. The modification made was to retain only those links where the "time last observed" field was the same as the day for that link file, e.g. 15th for the links file "links.20040115.gz".

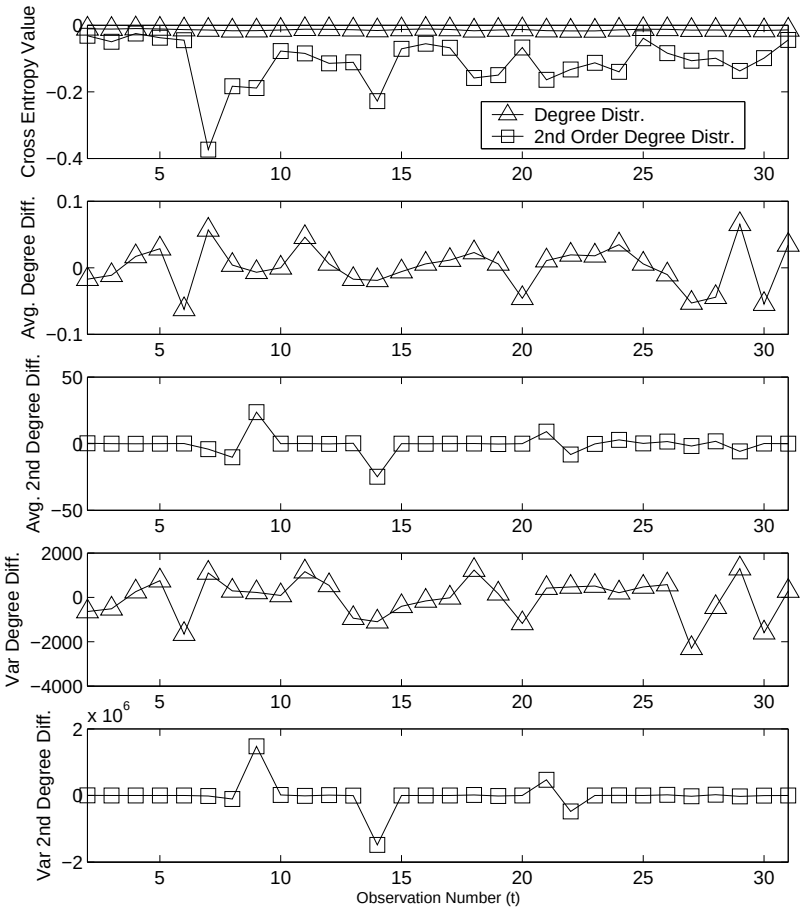


Fig. 4. Measurements of the time series of AS-level topology for January 2004.

Figure 4 was formed using the same procedure as used for the AS-graphs with a 3 month spacing. The cross entropy results are similar to that of Figure 2 in that the second order degree had more evident changes compared to the degree distribution. Significant changes in the second order degrees cross entropies at observations 7, 8, 9 and 14. The successive differences of average and variance of the second order degree distribution also indicate significant change at 9 and 14.

6 Conclusion

We propose a simple method for identifying change in a network's topology. We demonstrated empirically that the cross entropy between successive network observations can be used as a measure of change. As this is work in progress

there are areas requiring further study. These include; a comparison of the degree distribution approach with other measures of dynamic networks like that in [3]; a study of more networks, both large and small, and particularly networks that have the presence of a known topological change; and finally an investigation of the use of this method to compare parts of a single network (compare sub-graphs) and so perform a spatial comparison rather than a temporal one.

7 Acknowledgments

Thanks to Matt Roughan for pointing me in the direction of the work of Li et. al., to Carey Priebe for discussions regarding Rényi entropies, and Miro Kraetzl for discussions on measures of network dynamics. Also, thanks to the anonymous reviewers for their constructive comments.

References

1. R. Albert and A.-L. Barabási. Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.*, 74:247–97, 2002.
2. F. Chung and L. Lu. The Average Distance in a Random Graph with given Expected Degrees. *Internet Mathematics*, 1:91–113, 2003.
3. L. E. Diamond, M. E. Gaston, and Miro Kraetzl. An Observation of Power Law Distribution in Dynamic Networks. In *Proceedings of Information, Decision and Control Conference*, pages 101–105, 2002.
4. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *Proceedings of ACM SIGCOMM*, 1999.
5. M. Fayed, P. Krapivsky, J. W. Byers, M. Crovella, D. Finkel, and S. Redner. On the Emergence of Highly Variable Distributions in the Autonomous System Topology. *ACM SIGCOMM Computer Communications Review (CCR)*, 33(2):41–49, 2003.
6. B. Krishnamurthy, H. V. Madhyastha, and S. Venkatasubramanian. On Stationarity in Internet Measurements Through an Information-Theoretic Lens. In *Proc. 1st IEEE Workshop on Networking and Databases*, 2005.
7. S. Kullback and R. A. Leibler. On Informations and Sufficiency. *The Annals of Math. Stat.*, 22:79–86, 1951.
8. L. Li, D. Alderson, W. Willinger, and J. Doyle. A First-Principle Approach to Understanding the Internet's Router-Level Topology. In *Proceedings of ACM SIGCOMM*, pages 3–14, 2004.
9. D. Magoni and J. Pansiot. Analysis of the Autonomous System Network Topology. *ACM SIGCOMM Computer Communication Review*, 31(3):26–37, 2001.
10. M. E. J. Newman. Assortative Mixing in Networks. *Phys. Rev. Lett.*, 89:208701, 2002.
11. M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256, 2003.
12. A. Rényi. *Probability Theory.* North Holland, Amsterdam, 1970.