Disassortativity of computer networks

Patrick Rubin-Delanchy*[†], Niall M Adams[‡†] and Nicholas A Heard[‡†]
*Department of Statistics, University of Oxford
Email: delanchy@stats.ox.ac.uk
[‡]Department of Mathematics, Imperial College London
[†]Heilbronn Institute for Mathematical Research,University of Bristol

***Abstract*—Network data is ubiquitous in cyber-security applications. Accurately modelling such data allows discovery of anomalous edges, subgraphs or paths, and is key to many signature-free cyber-security analytics. We present a recurring property of graphs originating from cyber-security applications, often considered a 'corner case' in the main literature on network data analysis, that greatly affects the performance of standard 'off-the-shelf' techniques. This is the property that similarity, in terms of network behaviour, does not imply connectivity, and in fact the reverse is often true. We call this disassortivity. The phenomenon is illustrated using network flow data collected on an enterprise network. Improved procedures are proposed, that take explicit account of this property, for spectral analysis and link prediction.**

I. Introduction

There is growing evidence that statistical, data-oriented approaches to enterprise cyber-security can provide effective additional protection over traditional techniques [1], [2], [3], [4]. In such applications, data often have a network (or 'graph-like') structure, e.g. computers communicating on a corporate network [1], buyers and sellers in the underground economy [5], user authentication networks [2], and so on. Understanding the patterns of connectivity is key to developing many cyber-analytics for detecting, for example, nefarious network traversal and/or recognaissance behaviour [1]. As a result, statistical methodology for network data analysis is of great importance to cyber-security. However, a property that is ubiquitous in networks encountered in cyber-security applications, yet relatively rare elsewhere, is that nodes seem to organise into clusters such that connectivity *between* clusters is stronger than within. Although network modelling is a booming area of Statistics [6], [7], [8], and graph theory is otherwise relatively mature in fields such as mathematics, probability or computer science, many standard methods of analysis are inadequate because of this phenomenon.

To make matters concrete, consider a large corporate or institutional computer network. On this network, there are workstations, Domain Name servers (DNS), web servers, printers and much more. As a rule, printers do not communicate with each other; workstations rarely connect to other workstations; web servers do not themselves browse the web. In summary, nodes that are *similar*, e.g. in terms of their role on the network, their behaviour, or their connectivity patterns, are often relatively *unlikely* to connect. Drawing from biological terminology (e.g., [9]), we call this *disassortativity*.

This phenomenon has important consequences for the induced graph of connections. To illustrate, Figure 1 shows two graphs of communications between computers on the Los Alamos National Laboratory (LANL) network, observed over one minute on the left, and five minutes on the right. The graphs were constructed from the "network flow events" dataset [10], [11], by assigning each IP address to a node, and recording an edge if the corresponding two nodes are observed to communicate at least once over the specified time period. It is a common exploratory procedure to count the number of triangles in a graph in order to gauge clustering as, by transitivity, if similar nodes tended to communicate with each other, triangles would occur. Here, remarkably, *there are no triangles* in either graph.

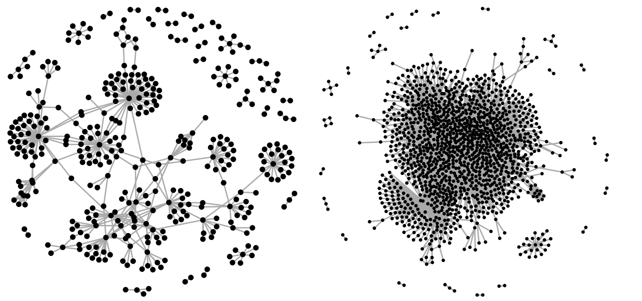

Fig. 1. LANL network flow graph. Left: first minute. Right: first five minutes

There are several reasons why disassortative behaviour might be observed in cyber-related data more generally. Networks are often structured into client-server relationships, where the clients are the instigators of communications, querying different servers for different services. In this model, client-to-client communications and server-to-server communications are more rare. Approximately bi-partite or k-partite network structure is also often induced by data collection mechanisms. Within a corporation, for example, most internal traffic can be recorded, as well as traffic between the inside and outside. However, obviously, outside-to-outside traffic is not observed, inducing partially bi-partite structure. Similarly, within the internal network, routers often do not record traffic between nodes on the same subnet, inducing k-partite structure. The superposition of different approximately bi-partite or k-partite network features results in a smoother continuum of network behaviours, but a strongly disassortative network structure remains.

Statistical literature on network data analysis is considerably more focussed on modelling assortative behaviour. In a seminal paper introducing ‘modularity’, Newman [12] writes “One issue that has received a considerable amount of attention is the detection and characterization of community structure in networks..., meaning the appearance of densely connected groups of vertices, with only sparser connections between groups”. “A tutorial on spectral clustering” [13], a reference for many data analysts and researchers worldwide (e.g. cited almost 4000 times according to Google Scholar, May 2016), considers only the first K eigenvectors of different types of Laplacian. As these can in each case be interpreted as solutions to relaxed min-cut problems [13] which, loosely speaking, seek to partition the nodes into densely connected clusters, they do not make good representations of (partly) disassortative networks. For this, the last eigenvectors must also be used, as we describe in the next sections.

II. Spectral analysis of network flow graphs

We will consider only undirected, simple graphs. The adjacency matrix of such a graph is denoted A, where A_{ij} is one if the nodes i and j share an edge, and zero otherwise. Hence, A is a symmetric $n \times n$ binary matrix, where n is the number of nodes. Its degree matrix is a diagonal matrix D where $D_{ii} = d_i = \sum_{j=1}^{n} A_{ij}$. Its normalized Laplacian is $L = I - D^{-1/2}AD^{-1/2}$ [13], where I is the identity matrix of order n. A common recommendation for representing the nodes of a graph as points in a space is to compute the first K eigenvectors of L, column-bind the vectors to form an $n \times K$ matrix, and take the rows of that matrix to represent the K-dimensional locations of the n nodes in space.

However, this is not entirely appropriate for graphs with a strong disassortative structure. To see why, consider that the eigenvectors $e^{(k)}, k = 1, \ldots, n$ of L minimise

$$\sum_{i,j} A_{ij} \left(\frac{e_i}{\sqrt{d_i}} - \frac{e_j}{\sqrt{d_j}} \right)^2 ,$$

under the constraint that the vector $e = (e_1, \ldots, e_n)^T$ satisfies $\|e\| = 1$. The actual minimum is achieved at the first eigenvector, $e^{(1)}$; the second eigenvector $e^{(2)}$ provides the next best solution that is orthogonal to $e^{(1)}$, and so on. The first K eigenvectors therefore embed the nodes into a space where nodes that are close are likely to connect.

For partially disassortative graphs, we have found the modified Laplacian $\tilde{L} = D^{-1/2}AD^{-1/2}$ [14] to have an easier interpretation. While the eigenvectors of L and $\tilde{L}$ are the same, the eigenvalues of $\tilde{L}$ are reversed and shifted down by one. These lie between -1 and 1, with negative values indicating disassortative network behaviour [14]. The sequence in decreasing order is hereafter referred to as the *graph spectrum*.

Figure 2 shows the spectra of different graphs generated from network flow data on the LANL network. Ten one-minute intervals were selected uniformly over the first day, and the graph of communications over each interval was constructed. Only the largest connected component of each graph was analysed. The spectrum of each is displayed as a line in the top panel. Because each has a different number of nodes ($n \approx 1000$), percentiles rather than raw indices for the sequences are used. Three representative network flow graphs are shown in the bottom panel.

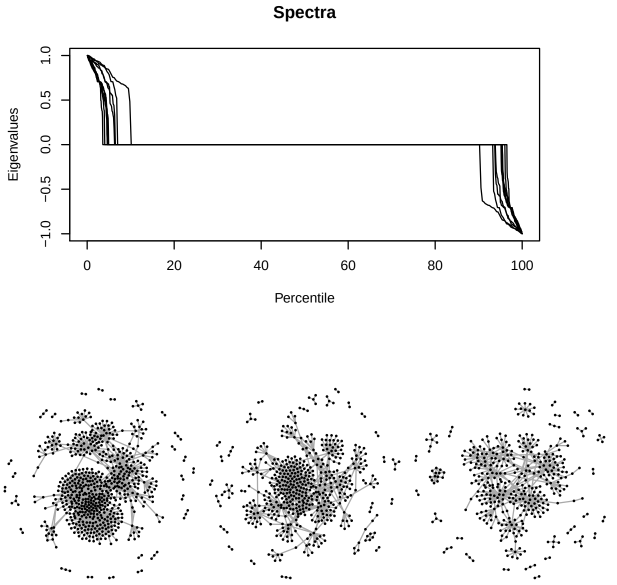

Fig. 2. Spectral analysis of LANL network flow graphs. The graph of communications between computers over a one-minute interval was constructed, for ten uniformly selected intervals in the first day. Top: the computed spectra. Bottom: three representative datasets (selected at random from the ten).

The spectra are distinctive. First, they are each almost perfectly anti-symmetric about 50%. This would not be expected for an arbitrary graph and in fact suggests that the network flow graphs are almost bipartite (although they are not exactly). More generally, the significant presence of negative eigenvalues is evidence of disassortative network behaviour. Second, a high number of eigenvalues are identically zero. This is due to a large number of nodes having *exactly the same* connectivity patterns. All of this information about the graph is ‘hidden away’ in the higher eigenvalues of L (zeros become ones and negative values now fall between 1 and 2), and would be lost if only the first K eigenvectors of L (e.g. $K = 10$) were computed, as is common for large graphs.

III. Application: link prediction

Discovering and explicitly handling the disassortative properties of computer networks is likely to improve the accuracy of network data analyses for cyber-security applications. We illustrate this with a link prediction example.

As a computer network is monitored in real time, it is natural to pay specific attention to the occurrence of new edges. They are, by definition, inconsistent with historical behaviour, and also indicative of nefarious activity such as network traversal or scanning [15]. On the other hand, over a large network such events occur regularly enough that it is useful, when shown a new edge, to be able to report its 'anomalousness' according to a predictive model. The value of such an approach depends strongly on the accuracy of the model. In a real application, a number of additional data sources could be brought to bear on this problem. However, here, for demonstration purposes, we use only graph information, and compare a link prediction algorithm that takes explicit account of the disassortative properties of the network to one that does not.

Given a graph of communications G, we consider two link prediction algorithms. The first, naïve, approach uses only the positive side of the spectrum, ignoring disassortative components. We compute the first K eigenvectors of $\tilde{L}$, corresponding to the highest eigenvalues, $\lambda_1, \ldots, \lambda_K$. The vectors are bound columnwise to form an $n \times K$ matrix, and node i is then represented by row i, denoted v_i. We form a diagonal matrix Λ containing the K highest eigenvalues in descending order. Finally, the probability of i and j sharing an edge is modelled as (a monotonic function of) $v_i \Lambda v_j^T$.

In the second approach, we use the first K eigenvectors corresponding to the highest K eigenvalues, *in magnitude*. In our data, this always results in using $K/2$ eigenvectors from the negative side of the spectrum, but this would not be true in general. We proceed as in the previous algorithm, so that each node i is represented by a row v_i', formed by column-binding eigenvectors from the positive and negative sides of the spectrum. This construction is theoretically justified since, for example, it generates consistent (i.e. asymptotically correct) clusterings under the stochastic block model [14]. The edge probability is modelled as $v_i' \Lambda' v_j'^T$, where Λ' is diagonal, containing the K eigenvalues with their original sign, in decreasing order of magnitude. In both models, $K = 10$.

For three one-minute intervals, uniformly selected over a day, we constructed network flow communication graphs from LANL data, as described previously, and fit both types of predictive models on each. For each interval, we then selected edges occurring within the next minute that were not in the original graph, but that did involve two nodes that had been active (so that each has at least one edge in the original graph). Every such edge is scored according to both predictive models. A similar number of pairs of nodes that did not communicate were selected, at random, as negative test-cases. In Figure 3, Receiver Operating Characteristic (ROC) curves are shown for the link prediction performance of both models, with each subfigure corresponding to one of the three time intervals. (For readers unfamiliar with this performance measure, the higher the curve, the higher the classification performance.) A false positive event is recorded whenever an edge is predicted but does not occur. A false negative event is recorded if an edge occurs when it was predicted not to. In all but the highest false positive regions (which are usually of lesser interest), the predictive model that uses both sides of the spectrum dominates.

IV. Conclusion

Graphs encountered in cyber-security applications, and especially network flow data, can exhibit strongly disassortative behaviour. We have shown that, because of this phenomenon, applying off-the-shelf statistical methodology to analyse such graphs can yield suboptimal results, both from the perspective of network characterisation and prediction. In a spectral approach, the solution is to consider both ends of the spectrum.

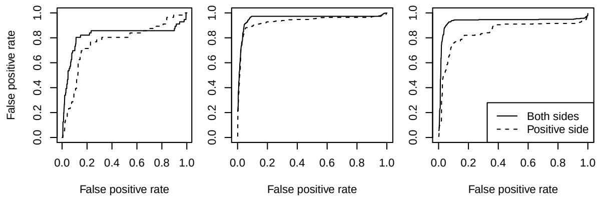

Fig. 3. Link prediction performance for new edges using only the positive side of the spectrum, versus both.

Acknowledgment

The authors would like to thank John M. Conroy (IDA Center for Computing Sciences, Bowie, USA) and Bill Oxbury (Heilbronn Institute for Mathematical Research, University of Bristol) for useful discussions and ideas.

References

[1] J. C. Neil, C. Hash, A. Brugh, M. Fisk, and C. B. Storlie, "Scan statistics for the online detection of locally anomalous subgraphs," *Technometrics*, vol. 55, no. 4, pp. 403–414, 2013.

[2] A. D. Kent, L. M. Liebrock, and J. C. Neil, "Authentication graphs: Analyzing user behavior within an enterprise network," *Computers & Security*, vol. 48, pp. 150–166, 2015.

[3] G. M. Jones and J. Stogoski, "Alternatives to signatures (alts)," *CERT Coordination Center, Software Engineering Institute*, vol. 20, p. 14, 2014.

[4] M. Morgan, J. Sexton, J. C. Neil, A. Ricciardi, and J. Theimer, "Network attacks and the data they affect," in *Dynamic Networks and Cyber-Security*. World Scientific, 2016.

[5] W. Li and H. Chen, "Identifying top sellers in underground economy using deep learning-based sentiment analysis," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*. IEEE, 2014, pp. 64–67.

[6] E. M. Airoldi, T. B. Costa, and S. H. Chan, "Stochastic blockmodel approximation of a graphon: Theory and consistent estimation," in *Advances in Neural Information Processing Systems*, 2013, pp. 692–700.

[7] S. C. Olhede and P. J. Wolfe, "Network histograms and universality of blockmodel approximation," *Proceedings of the National Academy of Sciences*, vol. 111, no. 41, pp. 14 722–14 727, 2014.

[8] C. Gao, Y. Lu, and H. H. Zhou, "Rate-optimal graphon estimation," *The Annals of Statistics*, vol. 43, no. 6, pp. 2624–2652, 2015.

[9] S. Khor, "Concurrency and network disassortativity," *Artificial life*, vol. 16, no. 3, pp. 225–232, 2010.

[10] A. D. Kent, "Comprehensive, Multi-Source Cyber-Security Events," Los Alamos National Laboratory, 2015.

[11] ——, "Cybersecurity data sources for dynamic network research," in *Dynamic Networks and Cyber-Security*. World Scientific, 2016.

[12] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[13] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[14] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, pp. 1878–1915, 2011.

[15] J. C. Neil, B. Uphoff, C. Hash, and C. Storlie, "Towards improved detection of attackers in computer networks: New edges, fast updating, and host agents," in *Resilient Control Systems (ISRCS), 2013 6th International Symposium on*. IEEE, 2013, pp. 218–224.