Characterization of Latent Social Networks Discovered through Computer Network Logs

Kevin M. Carter
MIT Lincoln Laboratory
244 Wood St
Lexington, MA 02420
kevin.carter@ll.mit.edu

Rajmonda S. Caceres
MIT Lincoln Laboratory
244 Wood St
Lexington, MA 02420
rajmonda.caceres@ll.mit.edu

Ben Priest*
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755
benjamin.w.priest.th@dartmouth.edu

Abstract

The collection of true and complete measurements for social network analysis is often limited and costly. Indirect measurements from electronic data sources have opened new avenues for extracting valuable information on social patterns. While the analysis of proximity data and email communication are a few data sources that have been heavily studied, the analysis of computer network logs and their robustness in capturing latent social behavior is an under-explored area. In this paper, we investigate the richness of network logs in identifying latent social networks of users within an enterprise. Specifically, we model the observable network artifacts on a per-user basis, and embed each user in graphs representing the enterprise. We analyze the structural properties of these graphs and discover that the latent social network shares many of the distinguishing characteristics of those observed in the real-world.

1 Introduction

Analysis of social networks has become a thriving area of research for identifying and characterizing social structures and behaviors of interest like homophily, self-organizing communities, and propagation of information. Often the data used to construct such networks comes from indirect, noisy measurements, with email communications [1], physical proximity [2] and web links [3] being a few examples. Therefore, it is of great interest to understand and quantify the robustness of these indirect measurements in capturing the true social relationships.

We focus on the extraction of social networks from computer network logs for the purposes of inferring information about an enterprise network. A truthful characterization of user behaviors and activities within a computer network is important with implications to network efficiency and cyber security [4, 5]. One may view the network as an overlay of the users' social network and the supporting computer network. Each layer imposes non-trivial constraints on the other, with the resultant model being more than the sum of the two. For example, different access privileges affect which physical edges are available for communication, while social relationships affect the intensity or traffic between computers. While recent work has demonstrated the use of network logs

*Work completed while employed at MIT Lincoln Laboratory

to characterize individual users and their local relationships [6], there has yet to be an analysis of how robust computer network logs are in capturing global characteristics of some unobserved social network.

In this paper, we address the robustness question by analyzing the properties of the latent networks constructed from computer network logs. We show that these latent social networks manifest many of the characteristics observed in real social networks such as sparsity, skewed degree distribution, high clustering coefficient, small world properties, and community structure [7]. This further implies that computer network logs can serve as reliable data proxies for capturing complex social patterns of users. This assessment has great implications to network efficiency and vulnerability [4], where the overlay of the human behavioral topology on the logical and/or physical computer networks presents a richer and more realistic view of the cyber network.

1.1 Related Work

Prior work has shown the ability to infer community structure within an enterprise from computer network logs [6]. We take a similar approach by considering topic similarity networks where topics arise from electronic footprints users leave as they utilize the computer network in their daily activities. While [6] took a metric-space approach, we model users as a social network and study the topological properties of the induced graph.

We leverage Latent Dirichlet Allocation (LDA) [8] to model our enterprise users. LDA has effectively been used to model user-created microtext [9, 10] and is a natural representation of computer network artifacts [11]. We extend that work by analyzing additional network artifacts, rather than just user search queries.

Our work is similar to that shown in [12], where the authors studied the topological properties of the Twitter network graph and contrasted them with information and social network traits. That work, however, leveraged observed relationships within an explicit social network – users directly tweeting each other – while our work focuses on latent relationships based on computer utilization.

2 Modeling Network Users

2.1 Data

We experiment on three different types of network artifacts: web search queries, web domains visited, and Kerberos resource authentications. We collected web proxy and Kerberos authentication logs indicating the activity of 3,715 users of a mid-sized enterprise network from September 1, 2011 to May 15, 2012, anonymizing the logs to protect user privacy. For ease of referral, we now use the term 'token' to represent an observation of a unique network artifact (e.g. search term, domain, or IP address).

To obtain the search query terms, we parsed the proxy HTTP request lines to derive a dictionary of 34,621 tokens. We parsed web domains from the proxy HTTP GET request lines to derive a dictionary of second-level domain name tokens. After filtering those domains which were visited by more than 50% of the user base – `google.com`, `twitter.com`, and `disqus.com` – or less than ten users (9,254 domains), the resultant dictionary contained 1,889 tokens. Kerberos resource authentication tickets are keyed by the IP address of the server hosting the resource in question. Hence, the 2,436 resultant tokens are IP addresses representing the host of the resource requested by a user.

We considered only those users with a minimum number of tokens, N_{min}, to remove inactive users. This value was $N_{min} = \{10, 50, 50\}$ resulting in {2654, 2782, 3052} users for Kerberos, search query, and web domain respectively.

2.2 Topic Modeling of Network Artifacts

Latent Dirichlet Allocation (LDA), first proposed by in [8], is a generative Bayesian topic model where observations are conditioned on latent multinomial variables with Dirichlet priors. For each choice of token type there are U users and W unique tokens, where each user u is associated with

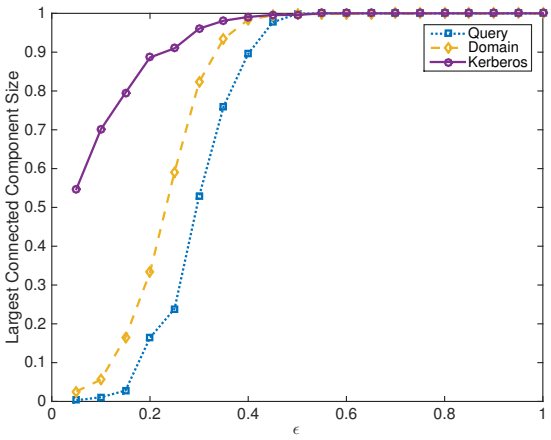


Figure 1: Size of the largest connected component – as a proportion of the total graph – over varying size ϵ-balls

a bag of N_u tokens of the specified type. Furthermore, we assume each u has a multinomial distribution Θ_u over K topics, and each topic k has a multinomial distribution ϕ_k over the W tokens; these distributions have Dirichlet priors with fixed hyper-parameters that are defined a priori. For each user u, LDA asserts u's bag of tokens is generated by sampling topic $z_i^{(u)}$ from Θ_u, and in turn sampling a token $w_i^{(u)}$ from $\phi_{z^{(u)}}$ for each $i = 1, \ldots, N_u$.

Training an LDA model to data entails estimating both the user-topic Θ_u and topic-token ϕ_k distributions from the observed users, given the fixed hyper-parameters of the Dirichlet priors and the number of topics K. We parameterize in order to optimize the *perplexity*, a measure of how well a trained model fits a held-out test set [8]. We select K at the point that increasing K produces diminishing returns. These values were $K = \{50, 100, 100\}$ for Kerberos, Query, and Domain respectively.

2.3 Defining Graph Structure

For each of our data sets, we infer an LDA model which implicitly projects each user u into a K-dimensional feature space, where u is represented by its vector of topic weights from its multinomial distribution over topics. A natural means of computing (dis)similarity between users is the cosine distance, which measures the length of the path between two points on the unit-sphere. Specifically, let us define the distance $d(u_i, u_j)$ between two users u_i and u_j as:

$$d(u_i, u_j) = 1 - \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|}, \tag{1}$$

where x_i is the mapping of user u_i to the topic model and $A \cdot B$ is the dot product, noting that $0 \leq d(\cdot, \cdot) \leq 1$.

To study the structure of latent relationships within the network, we embed users into a graph $G = (V, E)$ where $v \in V$ is the one-to-one mapping of users to graph vertices ($v_i := u_i$), and E is the set edges between vertices. For our purposes, we use undirected and unweighted graphs. We define G such that users are neighbors if they share usage patterns, or are 'close' in our metric space. Formally,

$$e_{ij} = \begin{cases} 1 & d(u_i, u_j) \leq \epsilon \\ 0 & d(u_i, u_j) > \epsilon \end{cases}, \tag{2}$$

for some value ϵ. Simply, we create an ϵ-ball around each user and draw edges between any users within the ball. Small values of ϵ require users to be highly similar in order to be neighbors, while larger values of ϵ relaxes this constraint. In Fig. 1 we plot the size of the largest connected component (LCC) within the resultant graph for various values of ϵ.

In many observed social networks, the largest connected component is a good representative of the whole graph. In all subsequent analyses, we define our graphs under analysis as those for which the LCC contains at least 90% of the graph. This results in a threshold of $\epsilon = \{0.25, 0.35, 0.45\}$ for Kerberos, Domain, and Query data sets, respectively.

3 Measures of Network Structure

Numerous metrics have been identified to capture distinguishing properties of social networks [7], and we detail several of them here.

Sparsity

The density p_G of graph G measures the portion of existing edges in the graph relative to the number of potential edges. Many empirical networks, but social networks in particular, are very sparse. That is, on average, vertices in the graph are only connected to a small constant number of other vertices.

Degree Distribution

The frequency distribution of vertex degrees of graph G – or simply its degree distribution – is another fundamental measure that has unique behavior when it comes to social networks. It is commonly characterized by most vertices having small degree and a few vertices with high degree, giving it a right-skewed shape.

Clustering Coefficient

The local clustering coefficient C_i is the proportional number of pairs in the neighborhood $\mathcal{N}_i$ of v_i, which are also neighbors themselves:

$$C_i = \frac{2|\{e_{ij} : v_j, v_k \in \mathcal{N}_i, e_{jk} \in E\}|}{|\mathcal{N}_i|(|\mathcal{N}_i| - 1)} \tag{3}$$

The global clustering coefficient C_G is defined as the average local clustering coefficient over all the vertices in graph G. The clustering coefficient measures the local cohesiveness of graph. Social networks are known to have high values of clustering coefficient relative to networks generated by random graph models such as Erdös-Rényi [13].

Average Shortest Path Length

The average shortest path L_G captures one notion of distance within graph G:

$$L_G = \frac{1}{n(n-1)} \sum_{i,j} l(v_i, v_j),, \tag{4}$$

where $l(v_i, v_j)$ is the shortest path between any two vertices v_i, v_j and $n = |V|$.

Small-World-ness

The small-world property was first defined by Watts and Strogatz [14]. A graph is considered "small-world" if it has a high clustering coefficient relative to the Erdös-Rényi graph, yet comparable average path length. We quantify the "small-world-ness" of G using the measure S_G defined in [15]:

$$S_G = \frac{C_G}{C_{rand}} / \frac{L_G}{L_{rand}}, \tag{5}$$

where L_{rand} and C_{rand} is the average path length and the clustering coefficient for the Erdös-Rényi graph. If $S_G > 1$, the graph is said be small-world.

Data Set	p_G	C_G	C_{rand}	L_G	L_{rand}	S_G	AUC
Query	0.014	0.52	0.014	4.00	2.64	23.74	0.982
Domain	0.012	0.76	0.012	5.58	2.71	31.19	0.993
Kerberos	0.104	0.89	0.104	3.29	1.90	4.93	0.998

Table 1: Measurements of topological properties within graphs constructed by user network artifacts

Edge Predictability

Social networks are known to exhibit transitive properties where nodes with common neighbors are likely neighbors themselves. The predictive power of the graph topology can be measured using the Adamic-Adar score, which computes the similarity between two vertices as the sum of the inverse log frequency of their common neighbors [3]. Higher scores suggest higher likelihood of an edge between those vertices. We measure the predictability of edges in graph G by randomly removing 10% of existing edges, resulting in $G = G_{train} + G_{test}$. We classify all missing edges in G_{train} by thresholding the Adamic-Adar score and using G_{test} as ground truth, and compute the area under the ROC curve (AUC) as done in [16]. This is then averaged over 10 trials. Higher values for this score suggests that G exhibits strong predictive power typically observed in real-world social networks.

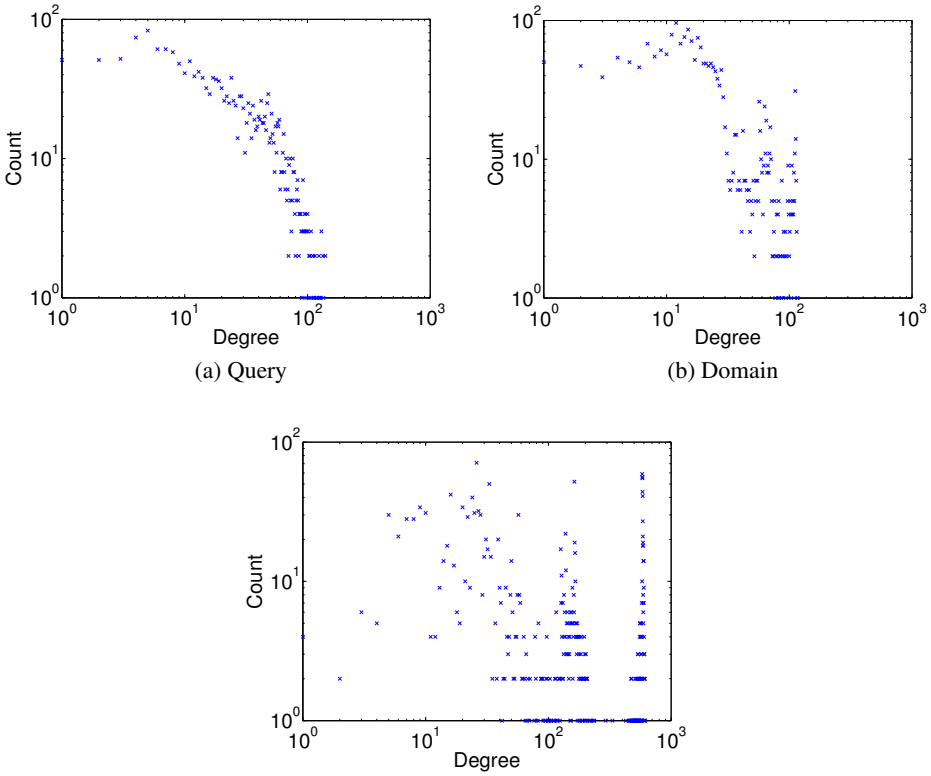
4 Results

In Table 1, we present for each of our latent social graphs the values of the metrics defined in Section 3. For comparison, we also show the corresponding values for the reference Erdös-Rényi graph with the same density. The latent social graphs exhibit density values representative of those observed in other social networks. In addition, there is significant predictive power of the graph topology, as the AUC when predicting edges was greater than 0.98 for all data sources. This is in line with results presented in [16].

Each of the studied graphs are characterized by high clustering coefficients and short path lengths, which lead to strong small-world properties as indicated by the high S_G values. It is important to note that while a larger value of S_G increases the confidence in the existence of the small-world phenomenon, this is not a metric. Hence, we make no claims that Query or Domain data exhibit stronger small-world effects than Kerberos, even though their S_G values are 5-7 times larger. What we do claim, however, is that each data set clearly exhibits small-world properties, which has implications to cyber security. In the case of Kerberos data, small-world properties have been associated with network efficiency and malware propagation [4, 5]. For the search query and web domain artifacts, small-world effects suggest collaborative environments but also broad attack surfaces for certain users. Understanding the interplay between the high degree nodes and the presence of small-world properties can help network operators with both planning and defense; optimizing load-balancing and identifying those users that are more prone to adversarial targeting.

In Fig. 2 we plot the degree distributions for the defined graphs. The degree distributions are skewed, more noticeably in the Query graph (Fig. 2a), but partially so in the Domain graph as well (Fig. 2b). Although the Kerberos dataset exhibits high clustering coefficient values and small-world properties, it has many more high degree vertices, representing those computer systems that are part of a common infrastructure across the user base.

Overall, our measurements imply that the inherent structure in the graphs defined by network artifacts is similar to that identified in real-world social networks, with the Query and Domain graphs being most similar. These results make sense as search and browsing are closely correlated with user interests, and social network forming mechanisms such as homophily and organization into communities are better captured through this type of network log [6]. The results demonstrate that network logs are robust in capturing the social layer in addition to their traditional use for characterizing computer traffic patterns. Therefore, they offer a versatile and unobtrusive data source that allows the simultaneous modeling and characterization of different layers that form the network.



(a) Query (b) Domain

(c) Kerberos

Figure 2: Degree distributions of G_ϵ derived from each network artifact (log-log scale)

5 Conclusion

In this paper, we presented a study of the latent social network structure of enterprise computer networks. Modeling users based on their observable network artifacts, we defined a graph for which users with similar utilization of the network were linked together. Our findings suggest that the structural properties of these graphs resemble those commonly found in observed social networks.

In future work, we wish to investigate how the joint analysis of such networks might reveal more robust social structures, as some network logs may be more reliable in capturing intrinsic behaviors of subsets of users. More representative latent social networks may be discovered by fusing information across different network logs. Network logs contain a wealth of information that is currently not being exploited, and this paper is just one step in learning the human aspect of computer networks.

6 Acknowledgements

This work is sponsored by the by Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

References

[1] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Physical Review E*, vol. 66, no. 3, pp. 2031–2034, 2002.

[2] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[3] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[4] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 65, no. 5 Pt 2, May 2002.

[5] I. Mishkovski, R. Kojchev, D. Trajanov, and L. Kocarev, "Vulnerability assessment of complex networks based on optimal flow measurements under intentional node and edge attacks." Springer Berlin Heidelberg, 2010, pp. 167–176.

[6] K. M. Carter, R. S. Caceres, and B. Priest, "Latent community discovery through enterprise user search query modeling," in *Proc. of the 37th Intl. ACM SIGIR Conference*, July 2014, pp. 871–874.

[7] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. of 36th Intl. ACM SIGIR Conference*, July 2013, pp. 889–892.

[10] B. Peng, Y. Wang, and J.-T. Sun, "Mining mobile usersÕ activities based on search query text and context," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2012, pp. 109–120.

[11] B. Priest and K. M. Carter, "Characterizing latent user interests on enterprise networks," in *Proc. of the 27th Intl. FLAIRS Conference*, May 2014.

[12] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: The structure of the twitter follow graph," in *Proc. of the 23rd International Conference on World Wide Web*, 2014, pp. 493–498.

[13] P. Erdös and A. Rényi, "On the evolution of random graphs," in *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 1960, pp. 17–61.

[14] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[15] M. D. Humphries and K. Gurney, "Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence," *PLoS ONE*, vol. 3, no. 4, Apr. 2008.

[16] Z. Liu, J.-L. He, K. Kapoor, and J. Srivastava, "Correlations between community structure and link formation in complex networks," *PloS one*, vol. 8, no. 9, 2013.