Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

Link Analysis of Incomplete Relationship Networks

Edward F. Harrington
Defence Science and Technology Organisation
Lock Bag 5076, Kingston, ACT 2604, Australia
Phone: +61-2-62650958, Fax: +61-2-62650925
email: edward.harrington@dsto.defence.gov.au

***Abstract*— We present a method of learning relationships at the triadic level of a relationship network. The method proposes learning linkages of a particular network using a Support Vector Machine (SVM) classifier trained on the known part of a relationship network. Using features drawn from the topological information of the two degrees of separation of a link a classifier learns whether two people of that link are related or not. We investigate empirically the performance of the technique for various relationship networks derived from email, web hyperlinks, and questionnaires.**

I. Introduction

Most social networks are either constructed from data gained from questionnaires or some form of information retrieval. One common problem is missing data. In the case of relationship networks this missing data is where we don't know if one *actor*[1] is linked to some other actor in the network. In automatic social network formation via web pages, as used in [3], inevitably there will be missing links simply because of incomplete sampling of the network. The dynamic nature of web pages and email also means we can't assume complete knowledge of linkages.

We wish to evaluate how effective the two degree of separation is at reconstructing relationship networks. Making the prime objective of this paper to be able to learn the relationships within the two degrees of separation of a network. The core of this method is to learn whether a link is missing or not based on the general behaviour of the two degrees of separation of that network. To learn whether a link is missing or not is dependent on network properties like the *transitive* nature of the three actors which make up the two degrees of separation. That is given three actors (A,B,C) they are transitive if that persons A is linked to person B and person B is linked to person C then A and C are linked. We propose using a pattern classifier to learn the nature of network properties of the two degrees of separation, properties like transitivity. For instance in friendship networks we expect the transitivity to be high because friends of friends are usually friends themselves. This is not necessary the case in an organisation's email system because the relationships can be rather hierarchical [6]. Suggesting other network measures as well as transitivity need to be considered when learning linkages. Measures like an actor's directional degree—the number of distinct linkages, either to or from that actor—are considered. The degree of a person in that relationship network is key to determining whether others in that network are likely to communicate with that person or not. In citation networks it has been shown that papers with a high *in-degree*[2] are more likely to exhibit *preferential attachment* [1] i.e. those papers with already high citations are more likely into the future attract more citations.

One application of learning network relationships is as part of a recommender system for web sites like: friendster `www.friendster.com`, a friendship building network; and LinkedIn `www.LinkedIn.com`, a business-oriented social networking site. These recommender systems would suggest to a person registered on these web sites other people who they are most likely to be interested in based on that person's network features.

Another application of having a more complete relationship network is it can be used as part of a system of email security. Viruses and spam present enormous problems to an organisation's email system. We propose that as part of a whole security system that incoming emails are checked to see if they might be from a person who is part of the recipient's relationship network. For instance when determining if an email is spam rather than just basing the decision on the content alone we propose considering whether that email is from someone from the recipient's relationship network drawn from the organisation's email system. By considering the relationship network combined with the spam content detection the number of legitimate emails being identified as spam can be reduced. This improvement is achieved by allocating more weight to an email whose content was detected as spam and the sender of the email was determined not to be part of the recipient's network. When dealing with viruses having knowledge to whether the email is from a person potentially known to the recipient or not, enables better ability to identify certain types of viruses. In some cases presents a mechanism to warn others in that recipient's network that they may have received a virus.

The rest of the paper is as follows. In Section II preliminaries are given. Section III provides details of the proposed

[1] An actor being the social unit used in a social network.

[2] In-degree refers to the number of edges or links into that node in the network.

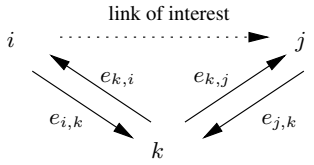


Fig. 1. The two degree separation for (i, j).

approach, and in Subsection III-A a description is given of the various attributes used to form the feature vector. We describe in Section IV the data sets considered in this study. The experimental results of the proposed method are presented in Section V, and Section VI concludes the paper.

II. Preliminaries

We model a network as a simple complete digraph $G(V, E)$ consisting of the set of all possible vertices V and edges $E = \{(i, j),\ \text{for } i \neq j \text{ and } i, j = 1, \ldots, N\}$, where given i and j are the vertices of the graph and each edge has a corresponding ordered pair (i, j).

A discrete scale is assigned to each ordered pair $(i, j) \in E$ denoted by $e_{i,j} \in \{0, 1\}$. We restrict for simplicity of discussion to a binary scale, i.e. if actor i says she or he is related somehow with actor j then the scale assigned is one, otherwise it is a zero indicating there is possibly no link between i and j. $H(V, E')$ is the graph (with no missing data) we are estimating given the graph $K(V, E'')$ with missing data, i.e. E'' E' E.

The two degree of separation is formed from a *triad* of two actors i and j and a third actor k (see Fig. 1). The third actor k is in the two degree of separation of (i, j) if it belongs to the subset of actors

$$\mathcal{K} = \{k \in V : R \text{ or } S \text{ or } T \text{ or } U\}, \tag{1}$$

with events $R = (e_{k,i} = 1 \text{ and } e_{j,k} = 1)$, $S = (e_{i,k} = 1 \text{ and } e_{j,k} = 1)$, $T = (e_{i,k} = 1 \text{ and } e_{j,k} = 1)$, and $U = (e_{k,i} = 1 \text{ and } e_{k,j} = 1)$.

III. Approach

For each edge of E a feature vector $\boldsymbol{x}$ and label $y \in \{\ \ 1, 1\}$ are assigned according to the scales associated with $K(V, E'')$. A label is $\ \ 1$ if the scale is 0 and 1 otherwise. See subsection III-A for details of how the features are constructed. Thus, the set of observations used in training of the classifier are $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{|E|}, y_{|E|})$ with indices allocated $(i \ \ 1)$ $N+(j \ \ i)$ for $i < j$ and $(i \ \ 1) \ \ N+(j \ \ i+1)$ for $i > j$. For example a network with $N = 10$ actors the link $(i = 3, j = 4)$ is allocated the feature $\boldsymbol{x}_{21}$ and corresponding label y_{21}, since $i < j$ the index is $(i \ \ 1) \ \ N + (j \ \ i) = 21$. The support vector machine (SVM) [8] was chosen as the classifier because of its good generalization in determining a link's label from its corresponding feature.

A. Feature vector

The attributes of $\boldsymbol{x}$ capture the significance of the triad (actors i, j and k) in relation to the link of interest. The directional information is partially captured in the attributes by including the degree's direction. The first six attributes are the in and out degrees (degree being the number links to that vertex) of the various actors of the triad (refer to Fig. 2). For example "in degree i" is $\sum_{j:i\neq j} e_{i,j}$ where $(i, j) \in E''$. The in in degree of an actor is often a measure of the actor's prestige. The higher the in degree the more prestigious the actor; and more likely they are to attract linkages with new actors, i.e. preferential attachment. Alternatively, the out degree of an actor is often a measure of the expansiveness of that actor to the rest of the network. When determining whether the link of interest is missing, or not, the direction of the degree is important, but so is the pairwise combinations of an actor's degrees. An example of why considering the pairwise combinations is important is that it is more likely that the link between actors i and j exists if there is a high out degree for actor i and a high in degree for actor j rather than if there is a high out degree for both actors i and j.

1. in degree i
2. out degree i
3. in degree j
4. out degree j
5. in degree k
6. out degree k
7. no. $e_{i,k}$ and $e_{j,k}$
8. no. $e_{k,i}$ and $e_{k,j}$
9. no. $e_{i,k}$ and $e_{k,j}$
10. no. $e_{k,i}$ and $e_{j,k}$

Fig. 2: Feature vector

To illustrate the in and out degree attributes consider the example network of Fig. 3. In this network of six actors the link of interest is from actor A to actor B. For this example the in and out degrees of $i = A$ are both two. When calculating the in degree of k it is different to in degrees of i and j. It is cumulative sum over all possible triads, so for the example in Fig. 3 $\sum_{k=\{C,D,E,F\}}$ in degree $k = 2 + 2 + 0 + 1 = 5$. For the out degree of k a cumulative sum is also used. Therefore, the out degree k in Fig. 3 is $\sum_{k=\{C,D,E,F\}}$ out degree $k = 1 + 1 + 2 + 1 = 5$.

The last four attributes are the number of two degrees of separation in a particular direction (see Fig. 2). Attribute 9, "no. $e_{i,k}$ and $e_{k,j}$" is the total number of two degrees of separation for the link between i and j, i.e. where $k \in \mathcal{K}$ and event U is true in (1). For example in Fig. 3 attribute 9 has the value of one, since only $k = C$ has $e_{i,k} = 1$ and $e_{k,j} = 1$. We use attribute 9 to measure whether the link of

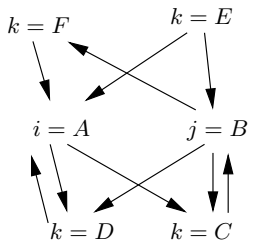


Fig. 3. Example network.

interest exists based on the transitive behavour of triads of that network. As mentioned previously, a triad is transitive if i is linked to k and k is linked to j then i is linked j. The theory of *cognitive balance* suggests in the case of friendship "like" networks that there is a tendency for relationships to be consistent, i.e. people tend to like their friends' friends [6]. Meaning if a network is transitive then attribute 9 will be high in value with a positive likelihood that the link exists between i and j. Attribute 10, "no. $e_{k,i}$ and $e_{j,k}$" is closely connected with the measure of *cyclicality* of triads within the network. Cyclicality means a triad has j linked to k, k is linked to i and i is linked to j. When the relationship is one of sharing resources then typically cyclicality in a network represents an indirect sharing of resources. For example in exchanging favors within a network when i does a favor for j, rather than j returning the favor directly to i, instead it is indirectly done, so j does a favor k then k does a favor i. In the case of the example of Fig. 3 the value of attribute 10 is two, as both actors D and F have $e_{k,i} = 1$ and $e_{j,k} = 1$.

While attributes 9 and 10 measure the positive likelihood that there exists a link from i to j actors, attribute 8 is more a measure of the negative likelihood case where it does not exist. Unconnected links between actors are sometimes referred to as *network holes*. In fact people take advantage of holes as a means to control of the flow of information between people not directly connected. These people are referred to as "brokers". Attribute 8 should measure the likelihood within a network that some actors in the network are brokers. For the example in Fig. 3 actor E is a "broker" for actors i and j, because it has $e_{k,i} = 1$ and $e_{k,j} = 1$.

Lastly, attribute 7 is capturing the case of whether i and j are linked based on they both are linked to the same person. Attribute 7 is closely related to the in degree of k but restricted to the triad. Therefore, attribute 7 crudely speaking is a measure of the importance of the third actor k within the triad. For the example in Fig. 3 this relative importance has the value of two, because both C and D have $e_{i,k} = 1$ and $e_{j,k} = 1$.

IV. Data sets

Enron's executive email network. This network consists of the emails sent between different executives of Enron from 1998 through 2002. We used a modified version of the data[3] used in [7]. We restricted the study to the 110 user ids who had sent more than 100 emails over the 189 weeks. If one or more emails were sent from user id i to destination user id j the value of one was allocated to $e_{i,j}$. When determining the destination user id no discrimination was made between "to", "cc' and "bcc", they were all treated as a link. The digraph produced had $|V| = 110$ and $|E'| = 1177$.

Krackhardt's High-tech Managers networks. The data used to construct the relationship networks consisted of questionnaire responses of 21 managers of a high tech manufacturing firm. Originally collected by Krackhardt [5] to study the managers perceptions of network structure. We consider only two of the relations studied: advice and friendship. The question used to study advice was "Who would you go to for advice at work (other managers only)". The question used to study friendship was "Who are your friends". Two relational networks were formed from the advice and friendship questions ($|V| = 21$, $|E'| = 190$) and ($|V| = 21$, $|E'| = 102$) respectively.

Freeman's Electronic Information Exchange System (EIES) network. The EIES network describes the acquaintanceships of 32 researchers at a conference in Sept. 1978 [4], [9]. The data used to construct the network was derived from a questionnaire. Five scales were used in the questionnaire: 4 = close personal friend; 3 = friend; 2 = person I've met; 1 = person I've heard of, but not met; and 0 = unknown name or non-response. The final digraph consisted of $|V| = 32$ and $|E'| = 759$.

WebKB networks. Two relationship networks were built using the hyperlinks contained in the home pages of students and facility members of the computer science departments of Washington and Wisconsin universities in 1996. The html home pages used were from `http://www-2.cs.cmu.edu/~webkb`. The Washington data produced a digraph with $|V| = 39$ and $|E'| = 55$. The Wisconsin data produced a smaller digraph with $|V| = 22$ and $|E'| = 29$. We created a directional edge between the home page person and a student or faculty member, if a hyperlink existed between their home pages. The intention was to use hyperlinks between home pages as a means of inferring relationships. To increase the likelihood that the triad existed we restricted the set of vertices to $C = \{i : \backslash(\sum_k e_{i,k} \quad 1$ and $\sum_k e_{k,i} \quad 1)\}$ making the set of edges $E' = \{(i,j) \in C \quad C : i \neq j\}$.

V. Experimental results

In all the experiments C-SVM with SVMlight optimization from the MATLAB library Spider [4] was used. To compensate for the fact that most of the networks had unbalanced classes the hyperparameter "`balanced_ridge`" was used. All the parameters were chosen using a small validation data set independent of the training data. A polynomial kernel of degree 3 was chosen for the SVM classifier. All the features had scale normalization. The main performance metric used was the *balanced error rate* (BER):

$$\text{BER} = 0.5 \quad (\text{FP} + \text{FN}), \tag{2}$$

where FP is false positive rate $= \frac{\text{number of false positives}}{\text{total number of negative instances}}$ and FN is false negative rate $= \frac{\text{number of false negatives}}{\text{total number of positive instances}}$. As a benchmark we compared the SVM results of the feature vector with a "naive" method which assumed transitivity. In this naive method the link between persons i and j was deemed to exist

[3] `http://cis.jhu.edu/~parky/Enron/enron.html`

[4] Library can be found at `http://www.kyb.tuebingen.mpg.de/bs/people/spider/download_frames.html`.

when there was no observed link between persons i and j only if person i was linked to person k and person k is linked to j.

Fig. 4 shows the results for the various data sets when linkages in networks were randomly removed to simulate missing data. The BER results were produced using 5-fold cross-validation. We used 5-fold cross-validation in an effort to give a good idea of the generalization performance of each method. The test labels were derived from the scales associated with $H(V, E')$, the graph with no links removed. Though the features used in testing were from the graph which had links randomly removed, i.e. $K(V, E'')$.

We see for all the data sets in Fig. 4 the SVM classifier can reconstruct links better than the naive method. In most cases of the networks in Fig. 4 as the number of missing links increased the balanced error rate differences between the two methods decreased. The only exception was the EIES friendship network where the methods BERs diverged. The most probably reason for this difference was that the EIES network has a low sparsity factor—ratio of the number links labeled as negative to the number labeled as positive—compared to the other networks, as displayed in Table I. One explanation for the divergence of the EIES BER results was due to the effective sparsity being increased by removing links, therefore decreasing the effectiveness of the naive method at predicting the positive labels. Looking at the results of Table II this seems to confirm this since the false negative rate for the naive method compared to the SVM is higher. When compared to the other networks the Enron BER result was the lowest. One possible reason for the lower BER result for Enron was due to the larger number of linkages used to train the classifier in the case of Enron compared to the other networks enabling, a better generalization of the SVM classifier.

In an effort to understand the results of Table II and Fig. 4 we used a Fisher linear discriminant [2] to determine the six highest ranking attributes in descending order of importance (see Table III). From the results of Table III we see attribute 9 was amongst the three highest ranked attributes. Highlighting attribute 9 as one of the key factors to the successful learning of links by both the SVM and naive methods. When looking at the Enron result we see in Table II that the naive method has a bias with a very low false negative rate (FN) but a high false positive rate (FP). The higher FP in the naive method was most likely due to the fact that for the Enron data transitivity in isolation was not sufficient to make the decision whether the link should exist or not. In an organisation like Enron it is fair to assume that there would be "brokers" wanting to build social capital and hierarchical relationships. The idea of brokers from a social capital perspective is they take advantage of unconnected individuals, network holes. By sharing information with them indirectly the broker increases their status in the network. Both hierarchical relationships and network holes are more likely to be captured by attributes 8 and 7. The lower FP rate of the SVM method for Enron could be explained by attributes 7 and 8. The higher FP rate of the naive method for the Enron network is not that well reflected in the BER measurement, and particularly when the network has a sparsity factor of 9.28.

TABLE I

SPARSITY FACTOR, NUMBER OF VERTICES $|V|$ AND EDGES $|E|$.

Network	$\|V\|$	$\|E\|$	Sparsity
Enron (email)	110	1177	9.28
Krackhardt (friendship)	21	102	3.32
Krackhardt (advice)	21	190	1.32
EIES (friendship)	32	759	0.35
WebKB (Washington)	39	55	26.65
WebKB (Wisconsin)	22	29	15.69

TABLE II

TEST RESULTS FOR 10 PERCENT OF LINKS MISSING.

Description		FN	FP	BER
Enron	naive	2.0(1.3)	26.4(0.6)	19.4(0.3)
	SVM	13.0(3.1)	8.1(3.1)	10.5(1.6)
Krack. frien.	naive	21.2(5.9)	50(2.3)	35.8(3.5)
	SVM	33.5(11.7)	17.2(2.4)	25.3(6.1)
Krack. advice	naive	3.9(3.2)	69.7(7.4)	36.8(2.6)
	SVM	24.0(8.5)	15.9(7.9)	19.9(6.4)
EIES	naive	34.6(3.6)	14.6(2.7)	24.6(2.8)
	SVM	27.9(3.5)	16.2(5.2)	22.1(4.1)
Webkb Wash.	naive	74.3(7.3)	1.5(0.8)	37.9(3.3)
	SVM	34.4(5.4)	11.7(3.8)	23.1(0.9)
Webkb Wisc.	naive	51.3(7.3)	3.7(2.1)	27.5(11.2)
	SVM	38.8(29.2)	4.6(2.1)	21.7(14.2)

TABLE III

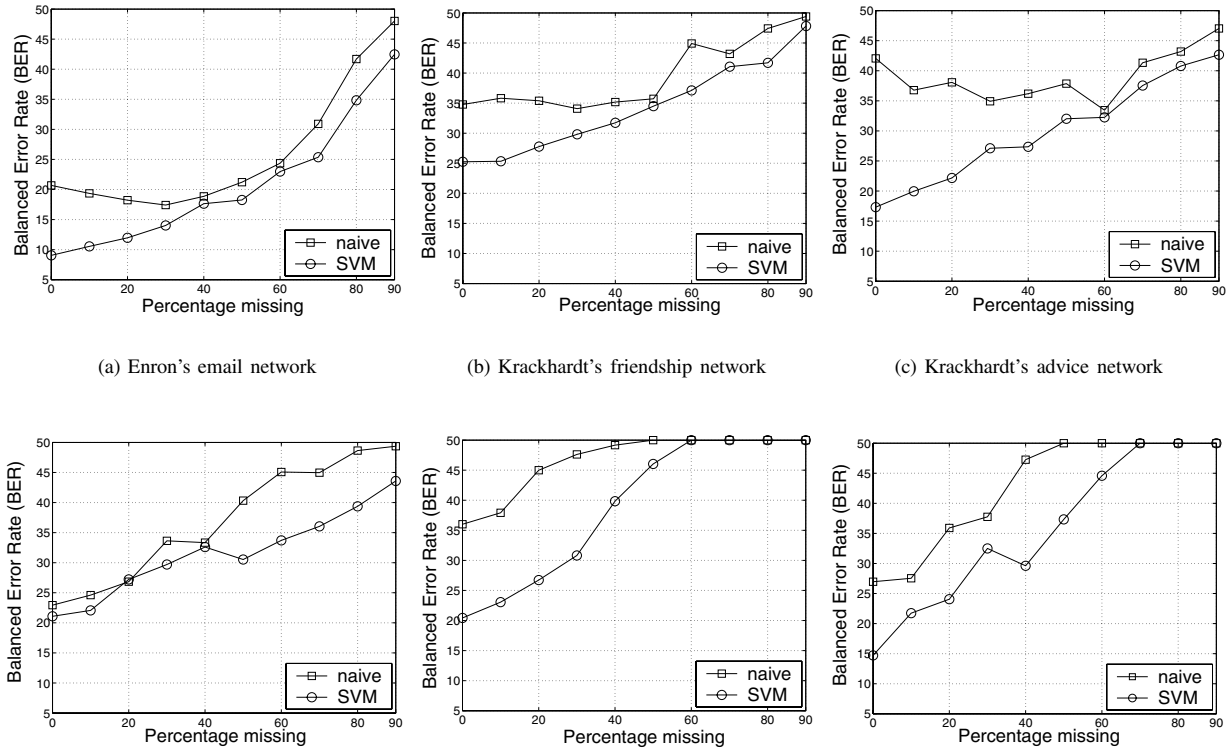
ATTRIBUTE RANKINGS USING FISHER DISCRIMINANT.

Network	1st	2nd	3rd	4th	5th	6th
Enron	9	8	7	5	6	10
EIES	9	8	7	10	6	5
WebKB (Washington)	3	2	9	5	1	6
WebKB (Wisconsin)	2	6	9	3	5	7
Krackhardt (friendship)	3	9	2	7	5	10
Krackhardt (advice)	3	9	2	7	5	10

For both the friendship and advice Krackhardt networks the BER was a fairer measure of performance compared to Enron due to the networks having sparsity factors closer to one.

We see from the naive method results of Table II that for the Krackhardt networks transitivity does not appear to be the most dominant network property. Krackhardt was interested in the perceptions of respondents of friendships and advice relationships of others within the network [9]. To evaluate these perceptions Krackhardt asked them to evaluate links between all actors, not just those actors the respondent was involved with. Krackhardt used a centrality measure of importance in the questionnaire finding that more important actors had better perceptions compared to other actors. This was supported by the results of Table III where the in degree of j (attribute 3) was ranked first. The in degree generally reflects that actor's prestige ,or importance, within the network. Interestly, the rankings of attributes for both advice and friendship networks were the same. This similarity between networks supports the hypothesis that the same network attributes at the triadic level were used to determine the perceived importance of a manager within Krackhardt's advice and friendship networks.

We now shift our focus to the WebKB results. Unlike the



(a) Enron's email network (b) Krackhardt's friendship network (c) Krackhardt's advice network

(d) Freeman's EIES network (e) WebKB's Washington network (f) WebKB's Wisconsin network

Fig. 4. Test BER percentages versus increasing percentages of missing links for naive and SVM approaches.

Krackhardt and Enron networks the naive method has a higher FN rate than FP rate (see Table II). This indicated that neither transitivity or cyclicality were major properties of the WebKB networks. From the results of the Table III we see that attribute 9 was only the third highest ranking and attribute 10 was not even in the top six ranking attributes. One explanation could be that the hyperlinks being predominately between students and staff members making transitive relationships less likely. Particularly in the case of the Wisconsin network the hyperlinks were most likely student homepages to their supervisors homepage explaining the higher influence of the out degree for the Wisconsin network (see Table III). In the Washington WebKB network certain graduate supervisors had higher in degrees making them more likely to be linked from their students homepages. This hypotheses was supported by the prominence of the in degree of j being the highest ranking attribute followed by the out degree of i. When looking at the BER results for WebkB the lower results compared to other networks are probably a result of the larger sparsity, and the smaller size of the WebKB networks.

VI. Conclusion

We presented machine learning methods to learn the relationships within the triad of a network. Through these machine learning methods we were able to deduce a better understanding of relationships at the triadic level. As well as being an effective tool for understanding triadic relationships we demonstrated empirically that the same methodology was effective at learning linkages when some links were missing.

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:247–97, 2002.

[2] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[3] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the Web. In *Proceedings of the 1st Conference on Email and Anti-Spam (CESA)*, 2004.

[4] L. C. Freeman and S. C. Freeman. A semi-visible college: structural effects of seven months of EIES participation by a social networks community. In *Electronic Communication:Technology and Impacts*, pages 77–85, 1980.

[5] D. Krackhardt. Cognitive social structures. *Social Networks*, 9:109–134, 1987.

[6] P. R. Monge and N. S. Contractor. *Theories of communication networks*. Oxford University Press, 2003.

[7] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan Statistics on Enron Graphs. *Computational and Mathematical Organization Theory*, 11(3):229–247, 2005.

[8] B. Schölkopf and A. Smola. *Learning with kernels*. Cambridge, MA:MIT Press, 2002.

[9] S. Wasserman and K. Faust. *Social network analysis methods and applications*. Cambridge University Press, 1994.