A Low Complexity Technique for Capture and Characterization of Social Network Topology

Sridhar Ramasamy[a], Randy Paffenroth[b], Anura P. Jayasumana[a]

[a]*Electrical & Computer Engineering Colorado State University Fort Collins, CO 80525*
[b]*Mathematical Sciences Data Science Program Worcester Polytechnic Institute Worcester, MA 01609*

Abstract

Social networks have millions of users and are complex in structure. Efficient techniques are required for measurement and characterization of these networks and extraction of embedded social relationships. This paper is an effort to capture the topology of a social network and predict the hop-distance between nodes. Techniques for capture of social network topology and distance relationships between node pairs are presented based on measurement of only a small subset of network paths, i.e., only a fraction of measurements normally required to fully characterize the network. Then low-rank matrix completion is used to extract the distance matrix from a subset of its entries.

Two methods are proposed. The first method uses a virtual coordinate system, consisting of hop distances from a set of anchors nodes to each of the nodes. This forms the virtual coordinate matrix. A small subset of this virtual coordinate matrix is used. The second method uses the hop distance between random pairs of network nodes. These random measurements of the distance matrix are further used in hop-distance prediction. Principle of low-rank matrix completion is then applied to predict the unknown measurements and hence the hop-distances between all the pairs of nodes.

Three different social networks (a Facebook sub-network, a collaboration sub-network and an Enron E-mail sub-network) are used to show the effectiveness of the approach. The size of network considered varies from 744 nodes to 4158 nodes. For both the methods, with fractional information available, the hop-distances are predicted and the results are compared with the actual distances. The results indicate that the connectivity information can be obtained with significant accuracy even with only 20% of the distance matrix elements.

Keywords: virtual coordinates, social networks, matrix completion

1. Introduction

Online social networking sites have a significant share of internet traffic and this share is growing enormously. Users of these sites tend to create networks of friends. The social network combined together becomes complex in structure. Let's take the example of Facebook which has millions of active users. The study on Facebook network on a global scale shows that it affirms the principle of "six degrees of separation" with almost 99.6% of all pairs of users within six degrees[21]. The average path length was found to 4.7 and Facebook is almost fully connected with 99.91% of individuals belong to a single large connected component. Interesting thing to note is that the Facebook graph as a whole is sparse, but graph neighborhoods possess dense structure[21]. Results from Flickr and yahoo! 360 shows that the average diameter of the graph slightly exceeds six degrees of separation. However with increasing edge density the diameter starts to fall[4]. This shows the complexity of network that we are dealing with and analyzing a social network could be voluminous.

A social network can be characterized by its adjacency matrix. A distance matrix can also be used to represent the network because one can be obtained from the other. The size of a social network graph is a big constraint and processing such a huge matrix would incur computation cost. In a social network, knowing the connectivity between nodes is not always possible owing to reasons such as privacy and secrecy. Also, measuring distances between all pairs of nodes is a nearly impossible task. The main contribution of this paper is introducing a technique to capture the social network topology from a much smaller set of known measurements of a distance matrix. In this paper we have used two different methods to obtain the topology of social networks with the proposed technique.

The first method used anchor-based Virtual Co-

ordinate System. This system uses M anchors and the distances from these anchors to all nodes in the network is known. Thus, a VCS is an M-dimensional abstraction of the network connectivity where M is number of anchors[9, 7, 6, 13, 5]. The placement and choosing of anchors is a key task. Choosing too many anchors will lead to increase in cost of VCS generation , whereas lesser number of anchors will lead to identical coordinates and local minima [8]. The current approach selects a random set of nodes from social network as anchors. Message is broadcast to all nodes inquiring about the connectivity and further forwarding of the messages is encouraged. The measurements that we get from responding nodes forms the subset of Virtual Coordinate Matrix. Then principles of low rank matrix completion is applied on this subset to capture the social network topology. The virtual coordinate matrix is a subset of distance matrix(explained in next section), and hence the subset of virtual coordinate matrix is also the subset of distance matrix. In that case, what if we have only random entries of a distance matrix? Thus, the second method uses pairwise hop distances between nodes to get back the topology of social network.

Section II discusses the background. Section III and IV discusses the algorithm and results for method based on virtual coordinate matrix. Section V and Section VI discuses the application of the technique with distance matrix. The paper is concluded with section VII.

2. Background

This section talks about the background on which this paper is built upon. Let's consider an undirected graph G with N nodes. This graph is characterized by its nodes and edges. The adjacency matrix gives the connectivity of the nodes in this graph. The Distance matrix D can be obtained from adjacency matrix. A distance matrix is composed of shortest hop-distances between any two nodes.

Our work is focused on social networks where, working with complete information of D is difficult because of the graph size. So a subset of D could be used effectively in the form of Virtual Coordinate(VC) matrix.

2.1. Virtual Coordinate Matrix

For an undirected graph, the distance matrix is symmetric. Let $D \in \mathbb{N}^{N\times N}$ be the distance matrix as shown in equation (1). The VC matrix is formed by selecting random anchors in the graph and computing the hop-distance to all the nodes i.e. it essentially means selecting those columns of entries from D that are chosen as anchors. Virtual Coordinate matrix is clearly a subset of the distance matrix D. Consider that M nodes are chosen as anchors. With an anchor-based Virtual Coordinate System (VCS), each node is characterized by a VC vector of length M. The inner matrix in equation (1), with N rows and M columns, forms the Virtual Coordinate matrix, P.

$$D = \begin{bmatrix} \begin{pmatrix} h_{n_1n_1} & \dots & h_{n_1n_M} \\ \vdots & \ddots & \vdots \\ h_{n_Mn_1} & \dots & h_{n_Mn_M} \\ \vdots & & \vdots \\ h_{n_Nn_1} & \dots & h_{n_Nn_M} \end{pmatrix} & \dots & h_{n_1n_N} \\ \vdots & \ddots & \vdots \\ h_{n_Mn_N} & \dots & h_{n_Mn_N} \\ \vdots & \ddots & \vdots \\ h_{n_Nn_N} & \dots & h_{n_Nn_N} \end{bmatrix} \quad (1)$$

where $h_{n_in_j}$ is the shortest hop distance between node n_i and node n_j. It is generally desirable to have only a small subset of nodes as anchors, i.e., $M \ll N$.

2.2. Low Rank Matrices

Singular Value Decomposition (SVD) [12] is widely used in principal component analysis (PCA)[1] and dimensionality reduction. SVD is used to analyze the low-rankness of the matrix. Let us consider a matrix A. SVD of matrix $A \in \mathbb{R}^{N\times M}$ can be written as

$$P = U\Sigma V^T$$

where $U \in \mathbb{R}^{N\times\min(N,M)}$, $\Sigma \in \mathbb{R}^{\min(N,M)\times\min(N,M)}$, $V \in \mathbb{R}^{M\times\min(N,M)}$ and $UU^T = VV^T = I$.

The diagonal entries of Σ are called the singular values of A. The rank of A is number of singular values not close to zero. The results section shows that the social network datasets are low-rank and also, from our prior work[10], we see that many real-world networks are low-rank.

3. Social Network Topology Capture from Partial VC Matrix

3.1. Capturing Connectivity Information

The Algorithm 1 gives a generic method to capture the connectivity from a social network.

This approach can be implemented on Facebook by broadcasting message to friends and in turn encouraging them to forward the message. Not all of them would be interested to cooperate and resulting measure will be a clear subset of VC matrix. A time to live t should be set to limit the messages in the network. As mentioned earlier, the social networks usually abide by the

Algorithm 1 Measuring Connectivity from Anchors

procedure AS A CENTRAL NODE
  M random nodes are selected as anchors
  for all Random nodes M **do**
    Send broadcast message to its neighbors with
    a TTL of t
  end for
  Response from anchors constitute subset of VC matrix.
end procedure

procedure AS AN INDIVIDUAL NODE
  if t==1 **then**
    Reply back to anchor with hop-count
  else
    Forward the message to neighbor with t decreased by one.
  end if
end procedure

"six degrees of separation" rule [20] ,and so it can be set to 6. It's assumed that random walks from anchors will lead to overlap of nodes. For a collaboration network, we can use this idea by looking at the co-authors of a paper and further branching out to look for other collaborators. For a E-mail network, The same method can applied by sending e-mails from one node to another with a limit in broadcasting.

3.2. Low-Rank Matrix Completion

This section discusses about the low-rank matrix completion. This closely follows the derivation in [16]. We provide this as it is essential for its application on social networks. The captured connectivity information forms a subset of P. The P is an incomplete matrix. Having found that these networks are low-rank, principle of low-rank matrix completion can be applied to this. We have formulated this from ideas in low-rank matrix completion[15, 2, 3, 18, 17]. The low-rank matrix completion can be explained as follows:

$$L = \arg\min_{L_0} \rho(L_0), \quad \text{s.t. } P_\Omega(P) = P_\Omega(L_0) \tag{2}$$

where ρ is the rank operator and Ω is the set of observed entries in P, so that P_Ω is a known subset of VC matrix P. We are trying to find a matrix L which predicts the unknown values in such a way that, $\rho(L)$ is minimized and it meets the constraints of existing hop-distance values. Equation (2) is an NP-hard optimization problem and can't be solved for large networks. But, recent results [15, 2, 3, 18, 17] allows us to address this issue by rephrasing Equation (2) to be a convex optimization problem

$$L = \arg\min_{L_0} \|L_0\|_*, \quad \text{s.t. } P_\Omega(M) = P_\Omega(L_0) \tag{3}$$

where $\|L_0\|_*$ is the nuclear norm, or sum of the singular values of L_0. This gives extends the solution to really large networks. This can be solved using splitting techniques and iterative matrix decomposition algorithms [15, 17, 18].

There are many standard libraries for solving these types of problems. For example, using the library CVXPY [11] in the scripting language Python [19] the optimization problem in (3) can be solved using code such as described in Algorithm 2

Algorithm 2 Matrix Completion

```
import cvxpy
L = cvxpy.Variable(N, M)
objective = cvxpy.Minimize(cvxpy.norm(L, 'nuc'))
constraints = [cvxpy.abs(D-L) <= epsilon]
prob = cvxpy.Problem(objective, constraints)
result=prob.solve()
```

4. Results - Virtual Coordinate Matrix

The results section analyzes the effectiveness of the proposed technique in recovering the VCs. Three networks have been used for this. A Facebook network with 744 nodes, A scientific collaboration network (Arxiv GR-QC (General Relativity and Quantum Cosmology)) with 4158 nodes, and An E-mail network from Enron with 3892 nodes. These datasets are available on Stanford Network Analysis Project [14]. The characteristics of the three network are give in Table 1. The histogram of the distance matrix is represented in Figure 1. The histogram gives us an idea of the distribution of hop-distances.

The first approach uses 20 anchors for Facebook network and 150 anchors each for other two networks. First the rank of the VC matrix is studied. Figure 2 shows the singular values on a natural log scale. This clearly shows that it is naturally low rank. The problems with singular values such as these are often treatable using matrix completion techniques. Next we randomly discard 5%, 10%, 20% up to 90% from the VC matrix. The matrix completion is applied and predicted matrix is obtained. This has been repeated three times to get an average result.

Network	Size	Diameter	Avg. length
Facebook	744	7	2.5549
Collaboration	4158	17	6.0479
E-mail	3892	5	3.139

Table 1: Characteristics of the Social Networks

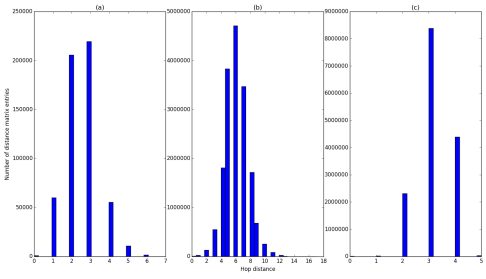

Figure 1: Histogram of the hop distances of distance matrix (a) Facebook network, (b) Collaboration network (c) Enron E-mail network

Two metrics have been introduced to show the accuracy of the proposed approach. We define mean error as follows:

$$E = \left[\sum_{i,j=1}^{N,M} |P_{ij}(f) - P_{ij}(0)|\right] / \left[\sum_{i,j=1}^{N,M} P_{ij}(0)\right] \quad (4)$$

where, $P_{ij}(f)$ refers to the VC matrix element denoted by row i and column j when f fraction of random anchor coordinates are missing. The mean error is a measure of percentage error in predicting the VC matrix. The Figure 3 shows the mean error with percentage of missing VCs.

The second metric proposed is absolute hop-distance error. This shows the deviation of hop distance from the complete VC matrix. The absolute hop-distance error is defined as:

$$H = \left[\sum_{i,j=1}^{N,M} |P_{ij}(f) - P_{ij}(0)|\right] / [n] \quad (5)$$

where $P_{ij}(f)$ refers to VC matrix element in i^{th} row, j^{th} column and f denotes the percentage of missing coordinates . n denotes the total number of elements in the VC matrix. The variation of absolute hop distance error with percentage of missing VCs is shown in Figure 4. Though, the mean error increases with a smaller number of known connections, the absolute hop-distance error gives an indication that hop-distance can be predicted with an error of 1 hop even when 90% of connections aren't available.

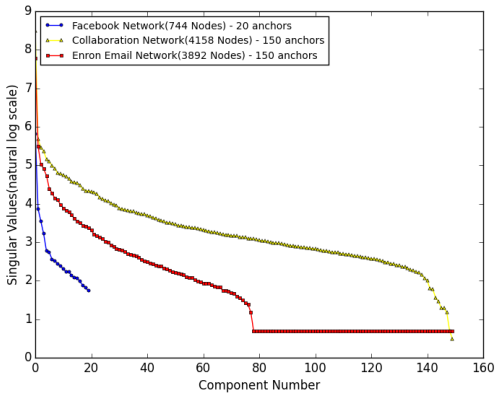

Figure 2: Singular values of VC matrix indicating that the adjacency matrices are naturally close to low-rank

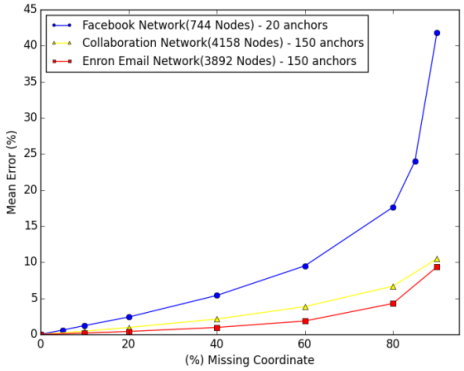

Figure 3: Mean error vs percentage of missing coordinates of VC matrix

5. Social Network Topology Capture from Partial Distance Matrix

The VC matrix is a random subset of the distance matrix corresponding to a few columns denoted by few anchors. Hence, a random subset of a VC matrix is almost the same as a random subset of the distance matrix. So we extended the approach to the distance matrix of network to study the performance. In case of an anchor based approach, we have a method to capture the partial information from the anchors which forms the subset of VC matrix. If the approach is not feasible, then based on available partial information between a pair of nodes, we can still apply the matrix completion and recover the topology of the network. A distance matrix is composed of shortest hop-distance between two nodes. Owing to the reason that the resultant matrix is a distance matrix of a graph, the measurement that we

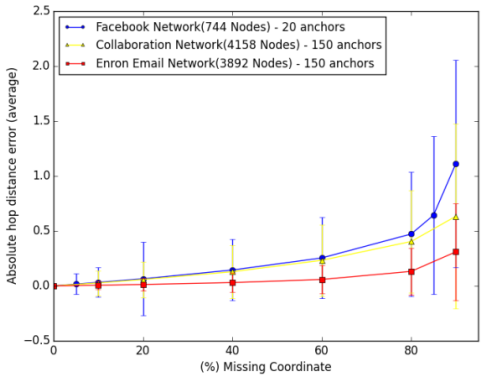

Figure 4: Absolute error in hop-distance with std. deviation vs percentage of missing coordinates of VC matrix

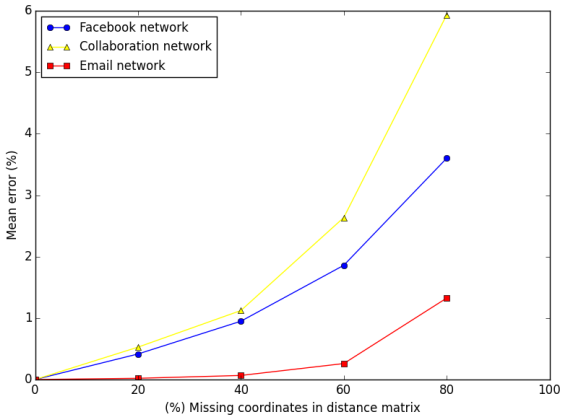

Figure 5: Mean error vs percentage of missing coordinates of distance matrix

have should also be shortest. This ensures that the adjacency matrix denotes accurate topology of network.

6. Results - Distance Matrix

This section analyzes the performance of the second method based on distance matrix.The same three networks have been used for this. The distance matrix is symmetric i.e. $D_{ij} = D_{ji}$. To remove one connectivity, both D_{ij} and D_{ji} should be dropped. Thus, entries from both lower and upper triangle of distance matrix should be dropped as a pair. The results have been obtained for 4 cases, 20%, 40%, 60%, 80% random drop of the distance matrix. The singular values of the distance matrix were observed to be low rank. The accuracy of this approach is evaluated the same way as evaluated for virtual coordinate matrix. In addition, the known entries of the original matrix are replaced in the predicted matrix. Also, the predicted matrix is rounded off to the closest integer value and finally the error is calculated. As mentioned 4, 4 is used to compute the mean error for the predicted distance matrix. The Figure 5 shows the mean error with percentage of missing distance matrix entries. Similarly, the absolute hop-distance error is calculated from 5 and the Figure 6 shows the absolute hop-distance error with standard deviation vs percentage of missing coordinates in distance matrix.

The histogram plot of absolute hop distance error for different percentage of entries removed indicates the error in terms of hop distance for the entries of matrix. Three histogram plot are included here. The Figure 7 shows the histogram plot for Facebook network with different percentage of distance matrix entries missing. This indicates that for this network, with 80% elements missing, the prediction is done well with an error of 1 hop. The maximum observed absolute hop distance error is 4, when 80% entries are removed.

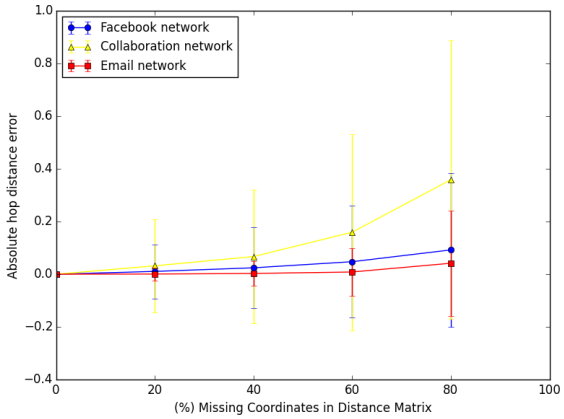

Figure 6: Absolute error in hop-distance with std. deviation vs percentage of missing coordinates of distance matrix

The Figure 7 shows the histogram plot for Collaboration network with different percentage of distance matrix entries missing. This indicates that for this network, with 80% elements missing, the prediction is reasonable well with an error of 2 hops. There are entries for other higher hop distance error values. The maximum observed absolute hop distance error is found to be 14 when 80% entries are removed.

Figure 7 shows the histogram plot for Enron E-mail network with different percentage of distance matrix entries missing. This indicates that for this network, with 80% elements missing, the prediction is done with an error of 1 hop. The histogram shows the result for range up to 2 hops, although the maximum obtained error is 3 hops

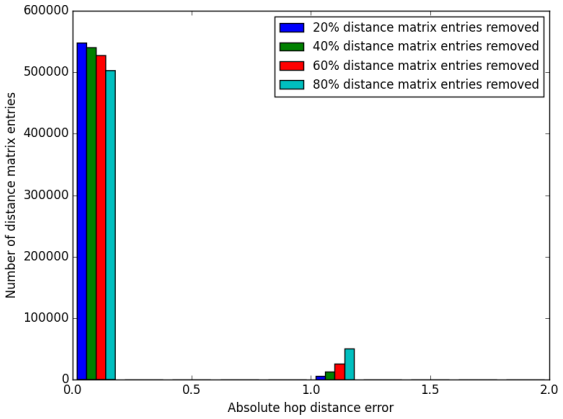

Figure 7: Histogram of the absolute hop distance error for different missing percentage of distance matrix for Facebook network

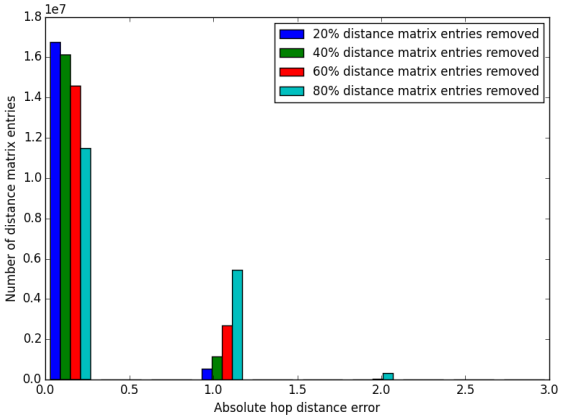

Figure 8: Histogram of the absolute hop distance error for different missing percentage of distance matrix for Collaboration network

for 80% entries removed.

There is an important point to note here. Looking at the Figure 1 and Table 1, it can be inferred that the average path length for the networks are different. The average path length of collaboration network is 6 and hop distances are predicted with an error of 2 for 80% elements missing. Similarly for Facebook network and E-mail network with average path length of 2.55, 3.139 respectively, the error in hop-distance is 1 hop. So we can infer that the error obtained is reasonable considering the average path length and the maximum hop distance.

7. Conclusion

A low complexity technique to capture social network topology was proposed in this paper. The proposed technique is evaluated with two approaches. The first being, capturing the topology

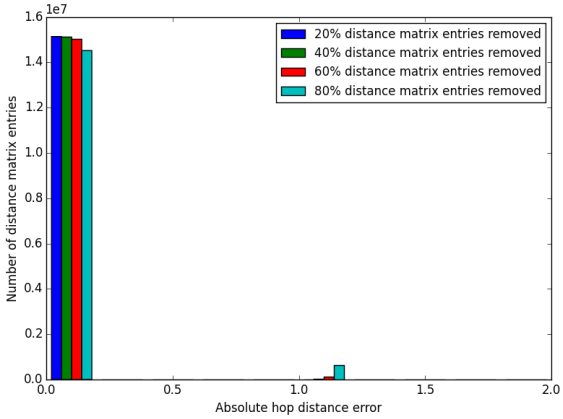

Figure 9: Histogram of the absolute hop distance error for different missing percentage of distance matrix for E-mail network

with fraction of information available in virtual coordinate matrix. The other uses available information between pairs of distances in the form of distance matrix. The second part, matrix completion is applied on the incomplete matrix to recover the complete topology of the network.

This technique of using virtual coordinates for graph opens a new method in graph compression. Also the possibility of obtaining the topology of a network with very less information is impressive owing to the reason it takes much lesser computation power and time. Many a times, obtaining information in the form virtual coordinate is not easy, and hence with even pairwise distance between nodes is sufficient to capture the social network topology.

[1] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York., 2006.

[2] E.J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):11, 2009.

[3] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009.

[4] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[5] F. R. K. Chung, L. Lu, , and V. Vu. Eigenvalues of random power law graphs. *Annals of Combinatorics*, 2003. p. 7:2133.

[6] F.R.K. Chung, Spectral Graph Theory, and American Mathematical Society.

[7] D.C. Dhanapala and A.P. Jayasumana. Dimension reduction of virtual coordinate systems in wireless sensor networks. In *Proc. IEEE GLOBECOM*, Dec, 2010.

[8] D.C. Dhanapala and A.P. Jayasumana. Convex subspace routing (CSR): Routing via anchor-based convex virtual subspaces in sensor networks. *Computer Communications. Jan. 2011*, 2011.

[9] D.C. Dhanapala and A.P. Jayasumana. Directional virtual coordinate system for wireless sensor net-

works. In *Proc. IEEE International Conference on Communications (ICC)*, 2011. June 2011.

[10] D.C Dhanapala and A.P. Jayasumana. Topology Preserving MapsExtracting Layout Maps of Wireless Sensor Networks From Virtual Coordinates. *IEEE/ACM Transaction on Networking*, 22(3):784–797, 2014.

[11] S. Diamond, E. Chu, and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization, version 0.2. 2014.

[12] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. ön power-law relationships of the internet topology. In *ïn SIGCOMM ́99: Proc. Conf. Applications, Technologies, Architectures, and Protocols for Computer Communication*, August 1999.

[14] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[15] Z. Lin, M. Chen, and Y. Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *arXiv:1009.5055*, page 23, 2013.

[16] R. C. Paffenroth, Anura P. Jayasumana, and Sridhar Ramasamy. Topology Maps and Distance Free Localization from Partial Virtual Coordinates for IoT. In *2016 IEEE Communications Conference*. IEEE, 2016.

[17] R. C. Paffenroth, P. Du Toit, R. Nong, L. L. Scharf, A. P. Jayasumana, and V. Bandara. Space-time signal processing for distributed pattern detection in sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):38–49, 2013.

[18] R.C. Paffenroth, R. Nong, and P. Du Toit. On covariance structure in noisy, big data. *SPIE Optical Engineering+ Applications*, pages 88570E–88570E, 2013.

[19] G. Van Rossum and F.L. Drake. Python Tutorial. *History*, 42:1–122, 2010.

[20] Jeffrey Travers, Stanley Milgram, Jeffrey Travers, and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[21] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.