Social Network Analysis of the Short Message Service

Vikrant Tomar, Himanshu Asnani, Abhay Karandikar, Vinay Chander, Swati Agrawal, Prateek Kapadia
TTSL-IITB Center for Excellence in Telecom (TICET),
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India
Email: vikrant@ee.iitb.ac.in, himasnani@gmail.com, karandi@ee.iitb.ac.in,
vinay87@gmail.com, agrawal.guddu@gmail.com, prateek@it.iitb.ac.in

***Abstract*—In this paper, we analyze patterns in the Short Message Service (SMS) behavior of customers in a large telecom service provider network. Toward this, we construct SMS Graphs, which are graphs induced by people exchanging SMSs, from the SMS Call Detail Records of the concerned service provider. These patterns are modeled by a weighted graph $G(V, E, W)$, in which the vertices represent the customers, and the edges and weights characterize the SMS transactions. We analyze properties of this graph, such as the distribution of component sizes, cliques, and vertex degrees. It is our belief that this study should enable the telecom operators to utilize the social behavior of their customers to design better service plans, and generate optimum incentives.**

***Index Terms*—Short Messaging Service, Social Network Analysis, Clustering**

I. Introduction

In recent years, Short Message Service (SMS) has emerged as one of the very popular means of communications. The trends indicate that the volume of SMS traffic is further expected to grow exponentially in the next few years. However, in spite of SMS volume increasing manifold, the operators' margins have not increased significantly, probably due to a mismatch between carrier competition and consumer expectation. The highly competitive mobile telecom market drives the service providers to continually rethink, and work out schemes that not only offer better incentives to their customers, but also maximize the revenue generated. This invokes a need of thorough analysis of the SMS behavior of the users, and realize recommendations to increase both utilization of network resources as well as revenue of the operators.

In this paper, we investigate how the social interaction of the customers can be utilized to achieve this goal. An SMS network could be visualized as an overlay of social network on the underlying telecommunication infrastructure. The degree of connectivity of users, base stations, mobile switching center (MSC) with other users, base stations and mobile switching center (MSC) is a reflection of social behavior. Hence, to better understand the traffic or rather network usage patterns, it is important to analyze this abstract social network. In this paper, we represent these social interactions in form of an SMS graph (graphs induced by people exchanging messages). Towards this, we analyze Call Detail Records (CDR), generated at the Short Message Service Center (SMSC), of SMS transactions from one of the largest telecom operators in India. To protect users' privacy the telecom operator had anonymized the data. Some information available in the CDR is originating User ID (an anonymous identification for a user), originating MSC, destination User ID, time of sending and delivering the message, and status of delivery. Some of the data statistics are listed in Table I.

In the recent times, world wide web or the Internet graphs or web-graphs has drawn the attention of plethora of researchers. Many of these researchers have shown that the nodes in a web-graph follow a Power Law for degree distributions [1], [2]. The Power Law suggests that probability, that a node has degree i, is proportional to $1/i^x$, for some $x > 1$, which is known as the power law exponent. Most studies on web-graphs has suggested the exponent to be 2.1 [2], [3]. There have been a number of studies on telecom graphs, as well. In [4], the authors have analyzed the properties of call and SMS graphs, and proposed the Treasure-Hunt model to describe the call graphs. Similar work can also be found in [5], [6]. Most of these studies have analyzed the graphs from web-graphs point of view, however, none has focused on the social networking aspects.

In this paper, we discuss the degree distribution of nodes, how clusters have emerged in form of various groups, and then we look at the strongly connected components and cliques. These connected components and cliques can be looked upon as indirect notions of social networking in the SMS graph. For the analysis, we have removed outliers, and seemingly toll-free and other service numbers from the data. By means of empirical analysis of the SMS traffic data, we establish that the nodes in an SMS graph also follow Power Law for degree distribution. Furthermore, connected components in an SMS graph also exhibit Power Law characteristics, this finding is in consonance with those on large web-graphs [2]. In the past, graph theoretic methods have been applied to web-graphs for ranking the results for web-search and browsing [7], [8], [9]. Our idea is to utilize the SMS graph, in conjunction with SMS traffic analysis, to design a tool which can generate recommendations for users based on their social behavior, as well as recommendations for the service providers to optimize their network and service plans to increase incentives for customers and revenue for the service providers.

II. System Architecture

In this section, we briefly describe the representative cellular system architecture followed by the SMS call flow. A simplified architecture of a cellular system has been depicted in Fig. 1. The concerned telecom operator's network consists of 100 Mobile Switching Centers (MSC) throughout the country. Each of these MSCs is connected to the two Signal Transfer

TABLE I
Some data about the studied telecom network

MSCs	$\approx$ 100
A2P SMSCs	2
P2P SMSCs	3
STPs	4
ITPs	4
Number of Cells	$\approx$ 8000
Average SMS Volume	$\approx$ 88 million/day
Voice/SMS Traffic	$\approx$ 2
Buffer Length at SMSC	0.6 million
Channels at SMSC	30 (E1 TDD Link)
Message Delivery Attempts (MDAs)	4000/sec

Points (STPs). The STPs are connected to IP Transfer Points (ITP) in a mesh topology by SS-7 links. The total traffic is equally divided between the two ITPs. The function of ITPs is to equally divide the incoming traffic onto outgoing links which gets connected to the SMSC. There are total five SMSCs in this network, however, only three are used for Person to Person (P2P) traffic, and the rest two are used for Application to Person (A2P) traffic. The two ITPs routes the incoming P2P traffic to the three P2P SMSCs in a round-robin fashion. SMSC operates on a store and forward mechanism, and is responsible for handling SMS messages.

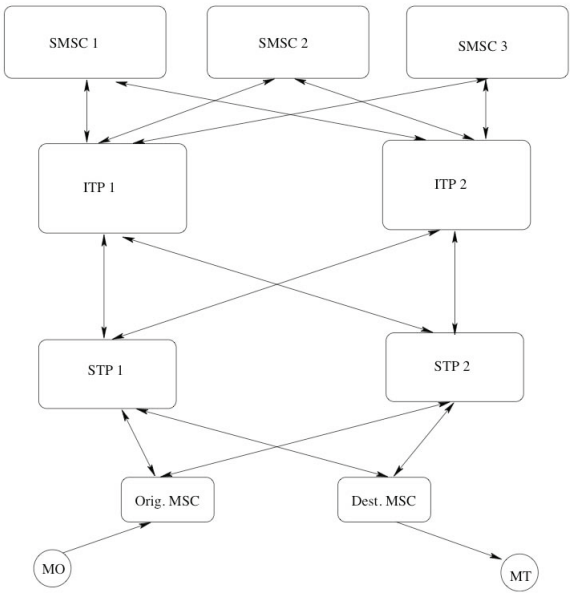

Fig. 1. **Simplified architecture of the discussed SMS Cellular Network**

When a user sends an SMS, it first goes to the base-station in the user's cell, and from there is forwarded to the MSC. The MSC forward messages as and when they arrive, and may queue them for systematic forwarding, however no processing is done at MSC. From MSC, the SMS reaches the SMSC, via STP-ITP mesh network. As an SMS arrives at the SMSC, it goes to the end of the arrival queue, and a Call Detail Record (CDR) is generated. While serving a particular SMS, the SMSC queries the HLR to locate the destination MSC and then delivers the message directly to the destination MSC. After this, it is the MSC's responsibility to deliver the message to the mobile station. After success delivery, another CDR is generated at the SMSC. If the HLR/VLR is not able to locate the mobile station then the SMS is queued in waiting. When the mobile comes online, this information is updated in HLR/VLR which then sends this information to SMSC, which in turn tries to deliver the SMS. Such pending SMSs will be queued for a maximum duration of one day, and then deleted.

III. SMS Social Graph

Using the CDR data, we have generated an SMS Social Graph for visualization of various clusters and groups. This sample SMS Graph, depicted in Fig. 2, highlights many small components. This graph has been generated using data of around 16 hours of Monday. In order to visualize the cluster formation, we created a base-user-pool with 10,000 users, randomly picked from the entire user-set, and then plotted the graph by taking only those transactions into account, in which atleast one of the users fell in the base-user-pool. In the graph, red edges represent lost messages. Such data can be used for network planning. For example, if many red edges correspond to SMS between a certain pair of MSCs, this pair can be investigated for performance bottlenecks. Furthermore, this graph may also be used to identify a large group of users, which can be used for tariff planning.

If we consider the users in the aforementioned graph to be nodes, and an SMS transaction as a link or edge connecting two nodes, we can construct a weighted graph $G(V, E, W)$, where V is the set of nodes, while E is the set of links. Total SMS transactions between node i and j is associated with a non negative weight $w_{ij} \in W$. Furthermore, depending on the hierarchy of social networking we are aiming, these nodes could be mobile stations, base stations or mobile switching centers. If the nodes are base stations or mobile switching centers, weights on the link indicate the fraction of total traffic generated at the transmitting node which is sent to the receiving node. Hence, other than non-negativity, weights satisfy additional constraints, i.e. $\sum_j w_{ij} = 1 \ \forall \ i$. Here, it is assumed that the traffic generated at each node is independent of the traffic received by the other node.

IV. Empirical Analysis of SMS Graphs

In this section, we analyze various structural properties of the SMS graphs.

A. Degree Distribution

The degree of a node in a graph is the number of connections or edges the node has to other nodes. Thus, a directed graph has two types of degrees viz., in-degree, and out-degree, which refer to number of incoming and outgoing links, respectively.

The *degree distribution* $P(d)$ of a graph is defined to be the fraction of nodes in the network with degree d. Thus if there are total n nodes in graph, and n_d of them have degree d, then $P(d) = n_d/n$ [1]. In addition to the in- and out-degree

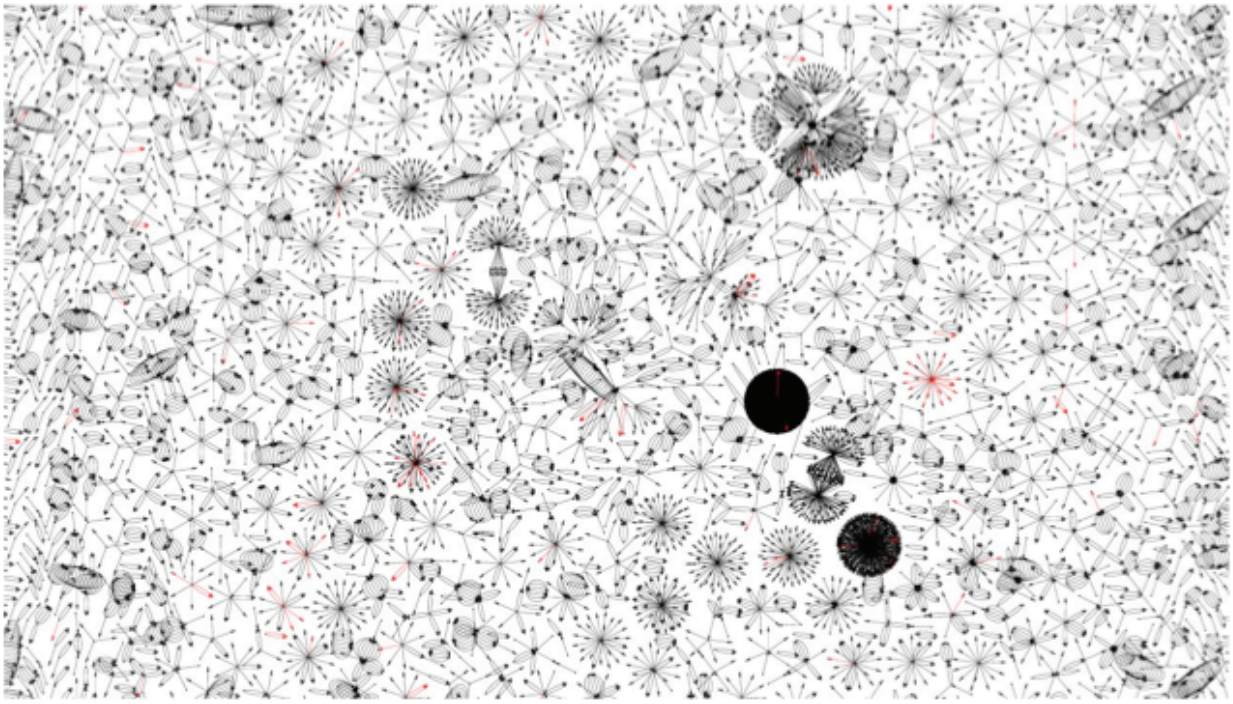

Fig. 2. **Visualization of Social Interaction in SMS transactions.**

classification, we can further associate two kinds of degrees to a node. First is the degree associated with the number of links to adjacent nodes (we refer to it as edge degree), and second is the degree associated with the total number of incoming or outgoing messages, which we call the message degree.

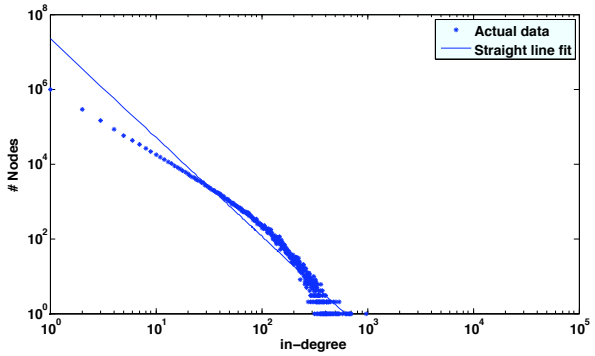

Fig. 3. **The message in-degree distribution.**

In Fig. 3 and Fig. 4, we report the behavior of the in-degree distributions for message in-degree, and edge in-degree respectively. All graphs are plotted in log-log scale with the degree on the X-axis, and the number of nodes/users on the Y-axis. The distribution in the figures demonstrates power law behavior with exponents 3.02 and 4.22 for message, and edge in-degree respectively. Furthermore, it can also be observed that the graph fits two intersecting straight lines better than a polynomial, meaning that there are two different regions of

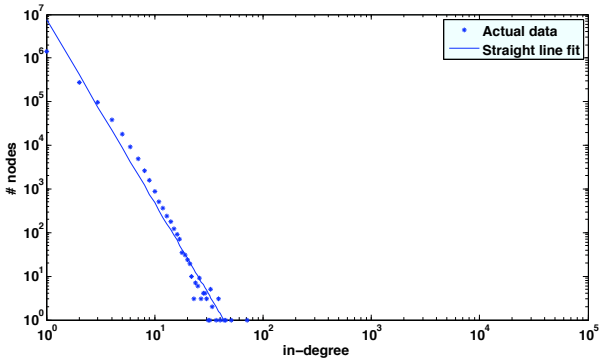

Fig. 4. **The edges in-degree distribution.**

linear growth.

Similarly, out-degree distributions are depicted in Fig. 5 and Fig. 4, for message and edge out-degree respectively. The power law exponents for the two graphs are 1.55 and 1.03. Note that there are a significant number of nodes having a very high out-degree. The users having very high values of out-degrees in terms of edges can be considered as spammers/outliers. We have refined the out-degree distribution by eliminating the outliers or spammers. For this purpose, any user sending messages to more than 100 users has been considered to be a spammer. The results are reported in Fig. 7 and Fig. 8. As expected, the removal of outliers results in better straight line fits. Also, the power law exponent has changed to 2.65 and 2.70, for message and edge out-degree, respectively.

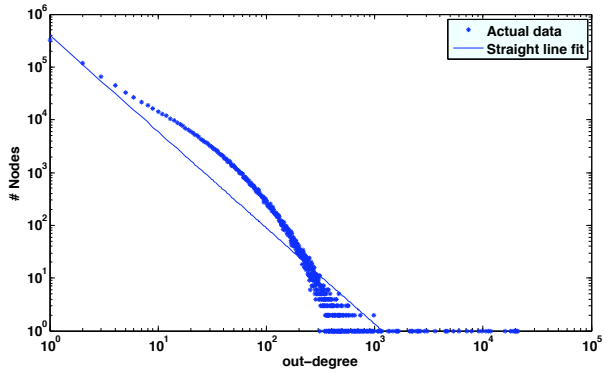

Fig. 5. **The message out-degree distribution.**

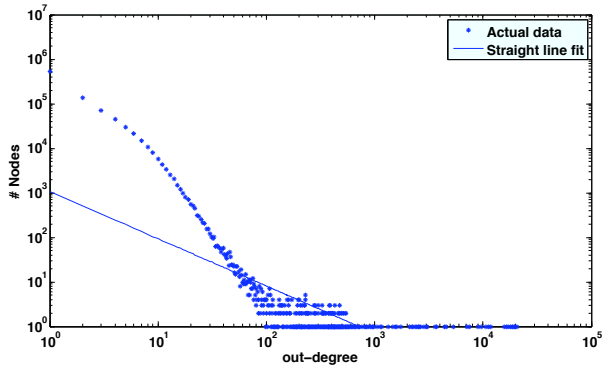

Fig. 6. **The edges out-degree distribution.**

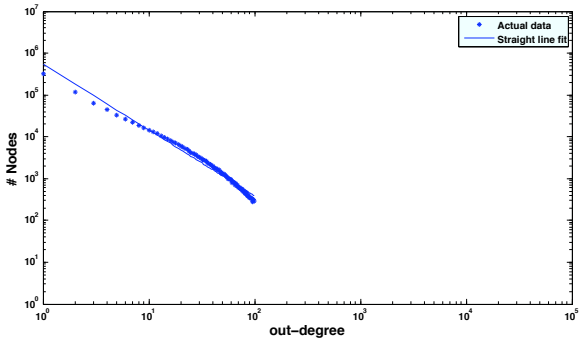

Fig. 7. **Out-degree distribution (messages) after the elimination of outliers.**

It can be deduced from the degree distributions that a very large percentage of users have small values of in/out-degree, and there are few users having very high out-degrees. Note that the values of power law exponents of the degree distributions in the SMS graph are close to 2.1, i.e, the power law exponent found in web-graphs [2], [3].

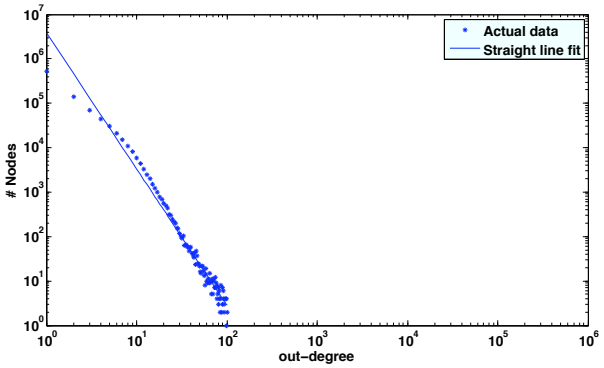

Fig. 8. **Out-degree distribution (edges) after the elimination of outliers.**

B. Clusters and Connected Components

Telecom service providers may use the information on clusters or connected components, including strongly connected components and cliques, to identify various user groups within an SMS graph/network. The size and number of such components gives extra insights about the right incentives to offer to such groups.

The weakly connected components or simply connected components are analyzed by converting the directed SMS graph into an undirected SMS graph. The distribution of weakly connected components in the network is depicted in Fig. 9. The general observation is that the number of components of smaller size is larger. From the figure, and data in Table II, following important observations can be made:

- 70% of the components are that of size two, which means that maximum user communication takes place in isolated pairs. This information can be used to design tariff plans providing discounts for SMSs between a particular user pair.
- The second important observation is that the number of users who communicate in isolated pairs is nearly equal to the number of users in the largest connected component, which may be worthy of further investigation.

TABLE II
DISTRIBUTION OF WEAKLY CONNECTED COMPONENTS

Total number of components	213042
Number of components of size 2	150267
Size of the largest component	472838
Value of power-law exponent	1.51

C. Strongly Connected Components

An Strongly Connected Component (SCC), in a directed graph, is defined as a sub-graph in which there exists a path between every pair of nodes [1]. We found most of the results of SCC to be very similar to those of connected components, except that the largest component was comparatively smaller. The distribution of SCC in the network is illustrated in Fig. 10. Furthermore, SCC in SMS Graphs exhibits similarity to SCC

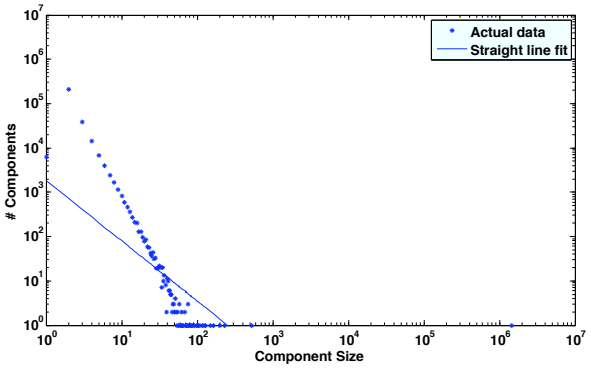

Fig. 9. **Distribution of weakly connected components.**

in web-graphs as in both cases, one giant strongly connected component exists. However, the largest SCC in the web graph covers a much greater percentage of the users as compared to that in the SMS graph. SCC, in web-graphs, have been used for designing web-search engines [7], [8]. Similarly, SCC in SMS graphs can be used to design a recommendation system by utilizing algorithms like PageRank, and HITS (Hypertext Induced Topic Search) [9].

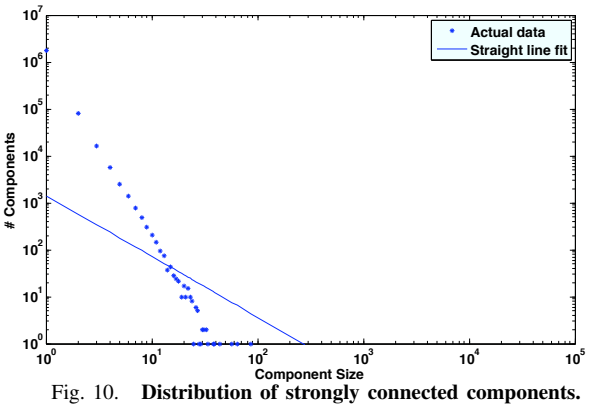

Fig. 10. **Distribution of strongly connected components.**

D. Cliques

A clique, in an undirected graph, is defined as a sub-graph in which there exists an edge between every pair of nodes [1]. From our analysis, we identified a number of cliques in the graphs, relevant data is provided in the table III. In our analysis, no clique of size greater than 6 was observed. Furthermore, the observation made after the removal of outliers was almost the same. There was slight decrease in the number of cliques of size 3 (difference of less than 30). The number of cliques of size 4, 5, 6, and 7 were found to be exactly same, thus it indicates that the spammers do not have a significant impact on the distribution of cliques.

V. Conclusions and Future Work

In this paper, we have analyzed various structural properties of the SMS graph, including the in-degree distribution, out-degree distribution, connected components and cliques. We

TABLE III

Clique size	Number of cliques
3	23602
4	1116
5	65
6	3

demonstrated that, similar to the nodes in a web-graph, degree distributions of nodes in an SMS network follow the Power Law. In addition, distributions of the connected components also exhibited the Power Law. Furthermore, in-degree and out-degree distribution, helped in identifying the spammers in the network. From the study on connected components, we show that almost 70% of the components are of size two, i.e., maximum user communication takes place in isolated pairs.

One of the possible uses of the degree distribution is to develop a traffic plan which benefits the users as well as the service providers. Information dissemination is another possible application of connected and strongly connected components. In a commercial network, the service providers would like to exploit the social networking aspects, and try to achieve maximum spread of information from minimum resources using underlying social dynamics. For example, two such models are Infect-Die Model for Dissemination [10], and Greedy Recommendation Scheme [11].

References

[1] A. Bonato, *A Course on Web Graphs*, ser. Graduate Studies in Mathematics. American Mathematical Society, 2008, vol. 89.

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the Web," *Computer Networks: The International Journal of Computer and Telecommunications Networking archive*, vol. 33, no. 1-6, pp. 309 – 320, June 2000.

[3] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for cyber communities," in *Proc. 8th WWW*, April 1999.

[4] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the structural properties of massive telecom call graphs: Fndings and implications," in *Proceedings of CIKM '06: the 15th ACM international conference on Information and knowledge management*, 2006, pp. 435–444.

[5] A. A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyzing the Structure and Evolution of Massive Telecom Graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 703–718, 2008.

[6] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec, "Mobile call graphs: beyond power-law and lognormal distributions," in *The 14th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, 2008, pp. 596–604.

[7] S. Brin, and L. Page, "The anatomy of a large scale hypertextual web search engine," in *Proc. 7th WWW*, 1998.

[8] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan, "Automatic resource compilation by analyzing hyperlink structure and associated text," in *Proc. 7th WWW*, 1998.

[9] A. N. Langville and C. D. Meyer, *Google PageRank and Beyond: The Science of Search Engine Rankings*. Princton University Press, 2006.

[10] P. T. Eugster, R. Guerraoui, A-M. Kermarrec, and L. Massoulieacute, "Epidemic Information Dissemination in Distributed Systems," *Computer*, vol. 37, no. 5, pp. 60–67, May 2004.

[11] P. Rojanavasu, P. Srinil, and O. Pinngern, "New Recommendation System Using Reinforcement Learning," *Special Issue of the Intl. J. Computer, the Internet and Management*, vol. 13, no. SP3, Nov 2005.