On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications

Amit A. Nanavati, Siva Gurumurthy, Gautam Das, Dipanjan Chakraborty, Koustuv Dasgupta, Sougata Mukherjea
IBM India Research Laboratory, New Delhi, India

Anupam Joshi*
Dept. of Computer Science, University of Maryland, Baltimore County, Maryland, USA
joshi@cs.umbc.edu

{namit,sgurumur,gautadas,cdipanjan,kdasgupta,smukherj}@in.ibm.com

ABSTRACT

With ever growing competition in telecommunications markets, operators have to increasingly rely on business intelligence to offer the right incentives to their customers. Toward this end, existing approaches have almost solely focussed on the individual behaviour of customers. Call graphs, that is, graphs induced by people calling each other, can allow telecom operators to better understand the interaction behaviour of their customers, and potentially provide major insights for designing effective incentives.

In this paper, we use the Call Detail Records of a mobile operator from four geographically disparate regions to construct call graphs, and analyse their structural properties. Our findings provide business insights and help devise strategies for Mobile Telecom operators. Another goal of this paper is to identify the shape of such graphs. In order to do so, we extend the well-known reachability analysis approach with some of our own techniques to reveal the shape of such massive graphs. Based on our analysis, we introduce the *Treasure-Hunt* model to describe the shape of mobile call graphs. The proposed techniques are general enough for analysing any large graph. Finally, how well the proposed model captures the shape of other mobile call graphs needs to be the subject of future studies.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data Mining*

General Terms

Algorithms, Experimentation

Keywords

Graph Analysis

*This work was done while the author was on sabbatical at IRL.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-433-2/06/0011 ...$5.00.

1. INTRODUCTION

As Mobile Telecom penetration is increasing, and even approaching saturation in many cases, the focus is shifting from customer acquisition to customer retention. It has been estimated that it is much cheaper to retain an existing customer than to acquire a new one [9, 12]. To maintain profitability, telecom service providers must control *churn*, the loss of subscribers who switch from one carrier to another. However, as the telecommunications markets grow more and more competitive, it is very easy for a consumer to churn because of low barriers to switching providers. In order to retain customers, the operators have to offer the right incentives, adopt the right marketing strategies, and place their network assets appropriately. To succeed in this goal, optimizing marketing expenditure and improved targeting are critical requirements.

Retrieving information from call graphs (where people are the nodes and calls are the edges) obtained from the Call Detail Records (CDRs) can provide major business insights to Mobile Telecom operators for designing effective strategies. A CDR contains various details pertaining to each call: when was it made, how long it lasted, who called whom, etc. Graph theoretic information from call graphs can allow service providers to better understand the underlying behavior of users, in a local as well as global context, in order to design incentives to increase subscriber loyalty and prevent/reduce churn. For example, if the call graph is disconnected into many small components then blanket advertising may be more appropriate as *word-of-mouth* spreading is impossible. Similarly, the presence of bipartite cores, which implies the presence of communities, can be supported with further analyses in order improve group targeting and retention.

Previously, a few experiments on call graphs of stationary telephone networks had been undertaken to determine parameters like cliques [2] and degree distributions [24]. However, to the best of our knowledge, this is the first study that attempts to discover and characterize a broad set of structural properties of mobile call graphs. The data we use comes from the CDRs of one of the largest Telecom operators in the world. In particular, we report findings on various topological properties of these massive call graphs, including degree distributions, strongly connected components, and bipartite cores. The presence of power law distributions is ubiquitous in many parameters of the call graph, a typical signature of its scale-free structure. Further, we observe interesting similarities and differences with respect to commonly studied networks like the WWW graph [20].

One of our primary motivations has been to characterize the shape of call graphs imposed by cellular phone users (referred to as mobile call graphs in this paper). For this, we employ reachability analysis, a technique that has been used to arrive at the Bow-Tie model for WWW graphs [7]. Having established a coarse structure using reachability analysis, we conduct additional experiments using novel techniques in order to reveal the finer structure of the graphs. An interesting revelation is that, whereas most existing graphs (hence their models) are based on the node distributions [7] in the components of the graph, our call graphs are better characterised by the *edge* distributions among the various components. We introduce the *Treasure-Hunt* model, an edge distribution based model, to characterise our mobile call graphs. The techniques proposed herein are general enough to be applied to the analysis of any network, and may be particularly relevant for social networks. In summary, the contributions of this paper are as follows:

- We study a broad set of parameters that reveal various structural properties of mobile call graphs.
- We describe novel techniques to determine the shape of large graphs
- We introduce the *Treasure-Hunt* model, an edge distribution based model, possibly the first topological model for mobile call graphs.
- We make a conscious effort to emphasize the practical implications of our findings in a way that can provide business insights and design strategies for mobile Telecom operators.

2. BACKGROUND AND RELATED WORK

Massive graphs originating from different sources like WWW and Biological networks have drawn the attention of plethora of researchers [20]. These graphs pose interesting challenges in terms of scalability, choice of parameters used to characterize them, and finally the techniques used for interpreting the graph structure. Even though many theoretical studies are available with several parameter sets, practical interpretation and utilization of those parameters (and results) are still lacking.

In the recent times, there is a lot of interest in studying World-wide Web and Internet graphs. Both [4] and [17] suggest that the *in* and *out* degrees of vertices on the Web graph exhibit power laws. Moreover, [4] has shown that most pairs of pages on the Web are separated by a handful of links, almost always under 20. This is viewed by some as a "small world" phenomenon. Determining groups of related pages in the WWW graph is another interesting problem. For example, [17] showed that *bipartite cores* in the Web graph represent implicitly defined communities. Our analysis reveals evidence of small world phenomenon in mobile call graphs.

A related area of research is the determination of the importance of pages (nodes) in the Web graph. The most well-known technique is *Page Rank* [6] which has been used very effectively to rank the results in Google search engine.

Another technique of finding the important pages in a WWW collection has been developed by Kleinberg [15] who defined two types of scores for Web pages which pertain to a certain topic: *authority* and *hub* scores. Documents with high Authority scores are authorities on a topic and therefore have many links pointing to them. On the other hand, documents with high hub scores are resource lists - they do not directly contain information about the topic, but rather point to many authoritative sites.

Yet another body of work has been undertaken in the determination topological model of the WWW hyperlink graph. Broder et al [7] showed that the Web has a *Bow-Tie* structure. This work outlines a general model but does not expose further details of the component structures. Examples of components are strongly connected component (SCC), incoming component (IN), outgoing links (OUT), and smaller components (TENDRILS). We mine finer structural properties of such components and identify parameters of interest to the telecom operator.

The Daisy model [11] is an attempt to further refine the WWW bow-tie model. Later, researchers have also tried to find out topological models for the Internet topology. The Jellyfish model [22] was one of the first in this direction. The Medusa model [8] is yet another model for the Internet topology, using a technique called k-core decomposition.

One of the first studies on call graph was performed on a graph of landline phones made on 1-day consisting of approximately 53 million nodes and 170 million edges [2]. The graph was found to be disconnected with 3.7 million separate components, most of them being pairs of telephones that called only each other. A giant component consisting of 80% of the total nodes was found. The diameter of this giant component was 20.

Aiello et. al. [24] experimented with the call graph of long-distance telephone traffic. The actual call graph showed that the degree sequence was not quite a perfect power law, and the authors introduced a unique class of random graphs with a power law degree sequence, called α-β graphs to capture the distribution. If $y(\gamma)$ is the number of vertices of degree γ, then the α-β graph is defined by the equation $y(\gamma) = e^{\alpha}/\gamma^{\beta}$. When the degree sequence of a power law graph is plotted on logarithmic scales, it forms a straight line. The parameters α and β specify this line; α is the point where the line intercepts the y axis, and β is the line's slope (rather the negative of the slope). Thus, α is the logarithm of the number of vertices of minimal degree, and β is the rate at which the logarithm of the number of vertices decreases as the degree increases. The best approximation has parameter values of roughly $\alpha = 17$ and $\beta = 2.1$. Values of β have implications on the structure of the graph.

Finally, in terms of business strategy design for telecommunication industry, many existing techniques exist based on mining of user profiles [12] as well as application of machine learning methods [9]. Most of these rely on the individual calling patterns of behaviors. We believe that structural findings from call graphs can augment and strengthen business intelligence directed towards the critical problems of customer targeting, campaign management and churn retention.

3. DATA SOURCES AND PREPROCESSING

A Call Detail Record (CDR) contains all the details pertaining to a call such as the time, duration, origin, destination, etc. of the call. The CDRs are collected at Base stations. Not surprisingly, a billion calls are made every month, and the data storage runs into terabytes. For our study, we analysed the intra-region calling patterns of four

Table 1: Details of data set used.

Region	Nodes	Edges	Period	Avg. Deg.	Type
$\mathcal{A}$-Region	224418	1816285	1 month	8.09	Dir.
$\mathcal{B}$-Region	1250656	4514528	1 week	3.6	Dir.
$\mathcal{C}$-Region	989573	4313797	1 week	4.35	Dir.
$\mathcal{D}$-Region	407332	1456645	1 month	3.57	Dir.

geographically and culturally disparate regions, for a single mobile operator.

A call graph G is a pair $\langle V(G), E(G))\rangle$, where $V(G)$ is a non-empty finite set of vertices (mobile users), and $E(G)$ is a finite set of vertex-pairs from $V(G)$ (mobile calls). If u and v are vertices of G, then an edge $\langle u, v\rangle$ is said to exist if u calls v. Hence v is adjacent to u in G.

To get the actual graph, we had to apply set of suitable filters and extract the best sample that would reflect the global calling pattern. The data set has the following characteristics:

- This study was done for a single mobile operator.
- The study was done for intra-region calls, and does not include long distance or international calls.
- For two of the regions, we collected all the calls made in a week, and for the other two, we collected all the calls made in a month. Interestingly, despite these durational and geographical differences, many parameters for these four regions are consistent with each other.
- Further, very short duration calls (less than 10 seconds) have been ignored as missed calls and wrong calls since they may yield incorrect results.
- Multiple calls between any two user or nodes is treated as a single edge. The resulting graph is *directed simple graph* with no self-loops or multiple edges.

For a given snapshot of the customer data records, G can be computed as follows. In a linear scan through all records, for each call, we write the corresponding ordered pair to a log file. At the end of the pass, the log is sorted and in a linear scan contiguous occurrences of each distinct ordered pair. Each ordered pair corresponds to an edge in G.

Table 1 shows the details of the data set used in this paper. While two call graphs have been generated for the span of 1 month ($\mathcal{A}$ and $\mathcal{D}$), the two are generated from call details records of 1 week ($\mathcal{B}$ and $\mathcal{C}$). Some basic graph properties such as the *number of nodes* n (also referred to as *graph size*) and the *number of links* m are reported. The *Average node degree* is defined as $\bar{k} = 2m/n$.

4. STRUCTURAL PROPERTIES OF CALL GRAPHS

In this section we analyse the structural properties of call graphs. Our analysis is based on a set of graph metrics that have been traditionally used to characterise large networks. In many cases, we use existing tools [21, 1, 10] for computing these parameters.

The basic properties of a network topology that characterize connectivity are degree related. The coarsest such property is the average node degree ($\bar{k} = 2m/n$), where $n = |V|$ and $m = |E|$ are the numbers of nodes and links respectively. Table 1 shows the average degrees of the call graphs.

4.1 Degree distributions

Distributions of degree gives information which average degree cannot, i.e. the number of nodes $n(d)$ of each degree d in the graph. We define this property as **node degree distribution** ($P(d) = n(d)/n$). The degree distribution $P(d)$ for directed networks splits in two separate functions, the in-degree distribution $P(d_{in})$ and the out-degree distribution $P(d_{out})$, which are measured separately as the probabilities of having d_{in} incoming links and d_{out} outgoing links, respectively. In Figure 1 and Figure 2, we report the behavior of the in-degree and out-degree distributions in log-log scale. We provided degree distribution results for regions $\mathcal{B}$ and $\mathcal{C}$ only because they are the largest ones and other two regions $\mathcal{A}$ and $\mathcal{D}$ were showing similar results.

Observing both in-degree and out-degree distributions, the call graph topology is found to be characterized by presence of a highly heterogeneous topology, with degree distributions characterized by wide variability and heavy tails. Observing log-log plots, we can see that degree distributions fit well to power law distributions. The power law exponent for all the regions are reported as the table representation in Figure 3. The results for other networks like Email and WWW are collected from [3]. While the power law exponent reported in many literature are for undirected graphs, our data set only consists directed graphs, which allows us to consider both γ_{in} and γ_{out} separately. In most real-world graphs, γ ranges between 1 and 4 (see [3] for a comprehensive list).

The in-degree distribution exhibits a heavy-tailed form approximated by a power-law behavior $P(d_{in}) \sim d_{in}^{-\gamma_{in}}$, and the value of the exponent of γ is between 2 and 3, very much like the WWW graph [7]. However, in the case of the out-degree distribution, the exponent is less than 2, very much unlike both the WWW as well as the Email graph. *The parameters of the four regions are rather close despite their geographic, cultural, and duration (1 week for two regions vs. 1 month for the other two) differences. The degree distributions imply that there are very few nodes that have very high in-degree or out-degree and therefore may be suitable for individual targeting by a telecom service provider.*

Another question of considerable interest in the study of networked systems is that of network resilience to the deletion of nodes. Suppose nodes are removed one by one from a network. How many must be removed before the giant component of the network is destroyed and network communication between distant nodes can no longer take place? Some networks, with highly skewed degree distributions, are found to be resilient to the random deletion of nodes but susceptible to the targeted deletion specifically of those nodes that have the highest degrees [5].

4.2 Degree correlation

The degree distribution is only one of the many statistics characterizing the structural and hierarchical ordering of a network; finding the degree correlations is another important parameter. Are the people who call a lot of different people also get called by a lot of different people (Is the in-degree of a node correlated with its out-degree) ? Are people with high in-degree talking to people with high in-degree ?

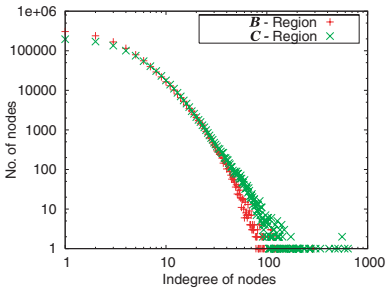


Figure 1: In-degree distribution (γ_{in} is 2.89961 and 2.89961)

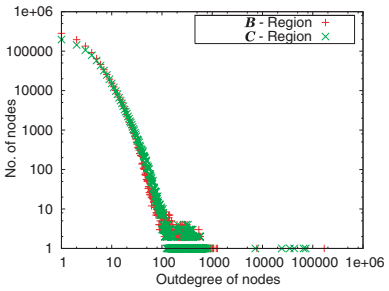


Figure 2: Out-degree distribution (γ_{out} is 1.70808 and 1.70808)

Results from our dataset (directed graphs)				
Region	$\mathcal{A}$-Reg.	$\mathcal{B}$-Reg.	$\mathcal{C}$-Reg.	$\mathcal{D}$-Reg.
In	2.76591	2.85924	2.89961	2.844
Out	1.50292	1.71374	1.70808	1.97345
Results from other datasets (directed graphs)				
Metric	Email	WWW	Software packages	
In	1.5	2.1	1.6	
Out	2.0	2.72	1.4	

Figure 3: Power law exponent values

In this section, we explore per node degree correlation as well as neighbour degree correlations in an attempt to answer such questions.

4.2.1 Single node IN-OUT degree correlations

First, we examine local one-point degree correlations for individual nodes, in order to understand if there is a relation between the number of incoming and outgoing neighbours for a single user. Since most of the analyzed degree distributions are heavy-tailed, fluctuations are extremely large so that the linear correlation coefficient is not well defined for such cases. We plot the average out-degree of nodes having the same in-degree.

Figure 4 shows the plots for the regions $\mathcal{B}$ and $\mathcal{C}$ (regions $\mathcal{A}$ and $\mathcal{D}$, not shown in the figure, have a similar correlation). A significant positive correlation between the in-degree and the out-degree of single nodes is found for both the sets. *This means that more popular nodes (whom a lot of people call) tend to point to more nodes (call a lot of people). More than 99% of total nodes have indegree less than or equal to* 100*; beyond which correlation disappears. The plots suggest that there is a correlation between the number of people calling a person and the number of people the person calls, up-to a point. Beyond a point, the correlation is absent. This suggests that up-to that point, people's popularity (incoming degree) is related with their "outgoingness" (outgoing degree). Beyond that point, the correlation falls, perhaps because, people with very high outgoing degree and low in-degree might be salesmen (who are seldom called back), and people with very high incoming and low outgoing might be (very) small businesses with advertised phone numbers, customer service numbers, or experts, for example.*

4.2.2 Neighbouring nodes IN-OUT degree correlations

A network is said to show *assortative* mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections [18]. Social networks (physics co-authorship, film actor collaborations, company directors) are often assortatively mixed, while technological (Internet, world wide web) and biological (food web, protein interactions) networks tend to be *disassortative*. In the case of the Internet, for example, it appears that the high degree nodes mostly represent connectivity providers, telephone companies and other communications carriers who typically have a large number of connections to clients who themselves have only a single connection.

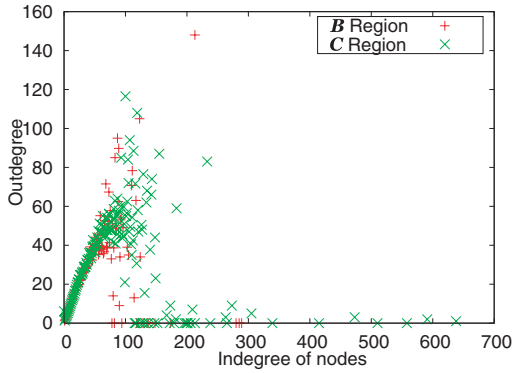


Figure 4: Correlation between in-degree and out-degree for regions $\mathcal{B}$ and $\mathcal{C}$

Table 2: Pearson's coefficient for the 4 regions

	$\mathcal{A}$-Region	$\mathcal{B}$-Region	$\mathcal{C}$-Region	$\mathcal{D}$-Region
In–In	0.293915	0.091079	0.175268	0.121410
In–Out	0.269043	0.311826	0.273044	0.036144
Out–In	-0.102487	-0.033359	-0.069215	-0.108047
Out–Out	-0.091477	-0.081860	-0.088832	-0.040468

The Pearson correlation coefficient (r) [18] denotes a single number that captures the assortativity of a network, and lies in the range $-1 \leqslant r \leqslant 1$. A negative value denotes disassortativity (-1 is perfect disassortativity), a positive value denotes assortativity (1 is perfect assortativity), and zero denotes no correlation.

For all the graphs (regions $\mathcal{A}$ through $\mathcal{D}$), we computed Pearson correlation coefficient to find out assortativity of the network. Table 2 provides the Pearson correlation coefficient for all the regions. The network is assortative for in-degrees and (relatively weakly) disassortative for out-degrees. The assortativity differs considerably with the zone.

As observed in [18], an assortative network helps in spreading and sustenance of an epidemic. Whereas an epidemic is expected to be restricted to a smaller segment of a population on a disassortative network. *In telecom networks, this property can be exploited for effective campaign management and spreading new services. For example, areas that exhibits high assortativity can rely on the word-of-mouth mechanism for spreading a new service. Whereas more advertising spend needs to be incurred for disassortative regions.*

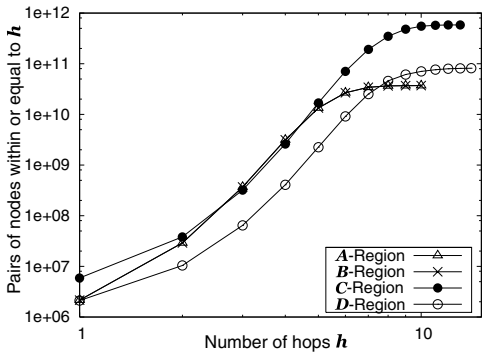


Figure 5: Neighbourhood function distributions (because the graphs have different effective diameters)

Further, assortativity has implications on the resilience of the network. It has been shown [19] that, for the most assortative network, with $r = 0.2$, it requires the removal of about ten times as many high-degree nodes to destroy the giant component when compared to the most disassortative one. *This once again implies that Telecom operators need to incur more expenses for retaining customers in disassortative regions.*

4.3 Neighbourhood distribution

Next, we compute neighbourhood function [21] for call graphs. It provides us ways to compare call graphs in terms of *hop-exponent*, distance distribution, and effective diameter.

The neighbourhood function, $N(h)$ for a graph also called hop-plot [13], is the number of pairs of nodes within a specified distance, for all distances h. The individual neighbourhood function for u at h is the number of nodes at distance h or less from u. It can be computed as follows:

$$IN(u,h) = |\{v : v \in V, dist(u,v) \leqslant h\}|.$$

The neighbourhood function $N(h)$ is the number of pairs of nodes within distance h, and is defined: $N(h) = \sum_{u \in V} IN(u,h)$. The plot of the neighbourhood function for all hop for all the regions are shown in the Figure 5. Also, $N(h) \propto h^{\mathcal{H}}$, where $\mathcal{H}$ is the $h$op exponent.

There are three interesting observations about the hop exponent that make it an appealing metric. First, if the power-law holds, the neighbourhood function will have a linear section with slope $\mathcal{H}$ when viewed in log-log scale. Second, the hop exponent is, informally, the intrinsic dimensionality of the graph. For example, a cycle has a hop exponent of 1 while a grid has a hop exponent of 2. Third, if two graphs have different hop exponents, there is no way that they could be structurally similar [21].

We compute the *hop-exponent* using linear fit on the neighbourhood function distribution reported in Figure 5. The slope of the linear fit called hop exponent are reported in the Table 3. For different circles of the telecom network graph; we consistently found hop exponent close to 4 and 5. The only other real-world graphs whose hop exponents we know are Int-11-97, Int-04-98, Int-12-98 and Rout-95 with hop exponents 4.62, 4.71, 4.86, 2.83 respectively [13]. *This suggests that our mobile telecom graphs are structurally as* dense *as those of the Internet graphs. Interestingly, even though the graph of regions $\mathcal{A}$ and $\mathcal{B}$ differ considerably in parameters*

Table 3: Hop exponent and effective diameter values (C is the constant for the linear equation fit in log-log scale for the equation $N(h) \propto h^{\mathcal{H}}$)

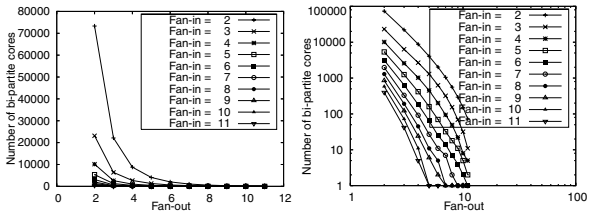
Region	$\mathcal{H}$	C	δ_{eff}
$\mathcal{A}$	4.53	14	8.10177
$\mathcal{B}$	4.53	14	13.9314
$\mathcal{C}$	5.52	14	8.07642
$\mathcal{D}$	4.94	13	8.93613

such as average degree, number of nodes and edges (Table 1), their hop exponents are similar (5).

Effective diameter gives us another parameter for effective measurement of the compactness of the network. For a call graph of N nodes with E edges, we can compute effective diameter based upon the equation [13]:

$$\delta_{\text{eff}} = \left(\frac{N^2}{N + 2E}\right)^{1/\mathcal{H}}$$

The effective diameter of a network is δ_{eff} if any two nodes are within δ_{eff} hops from each other *with a high probability.* The effective diameter for all the four regions are given in the Table 3. Our results indicates that most of the pairs of nodes are within 8 to 12 hops from each other. This in turn provides evidence of small-world phenomenon in mobile call graphs. We believe that this phenomenon can be further exploited to identify (social) communities.



(a) Standard scale (b) Log-log scale

Figure 6: Distribution of bipartite cores for the $\mathcal{B}$-region

4.4 Bipartite cores

To study whether there are communities of users within the call graphs, we find out possible presence of bipartite cores using the trawling technique [17]. A complete bipartite clique $\mathcal{K}_{i,j}$ is a graph where every one of i nodes has an edge directed to each of the j nodes. A bipartite core $\mathcal{C}_{i,j}$ is a graph on $i + j$ nodes that contains at least one $\mathcal{K}_{i,j}$ as a subgraph. Figure 6 (a) shows the results for the bipartite cores for the largest region $\mathcal{B}$. Similar distributions were obtained for the other three regions and have been omitted due to lack of space. For a bipartite core of $i + j$ nodes, fan-in (i) represents the number of nodes having edges to the j nodes. Fan-out (j) represents the number of nodes having edges from the i nodes. We observed a large number of small bipartite cores (e.g. $\mathcal{C}_{2,2}$). The largest bipartite core has $i = 11$ and $j = 6$. Figure 6 (b) shows the results for the bipartite core distribution in log-log scale. We observe that the distribution of bipartite cores closely follows the power law. In terms of the presence of bipartite cores, the call graph exhibits similarity to WWW [17] and web communities in blogs [16]. *These bipartite cores possibly*

represent user communities such as a community of patients and doctors. Hence, telecom service providers can identify these communities and target them with better incentives for retention.

4.5 Cliques

For mobile telecom providers, an (undirected) clique is useful for defining *closed user groups* (as they are commonly called), where discounts are given for all calls made within the closed user group. The number and sizes of such groups also gives us an idea of what are the right incentives to offer. The distribution of the clique sizes are shown in Figure 7. No clique of size larger than 11 is observed. Relatively, large number of cliques of size 3 is observed. Intuitively, this might be because it is rare for any two people who know each other not to know at least one common person.

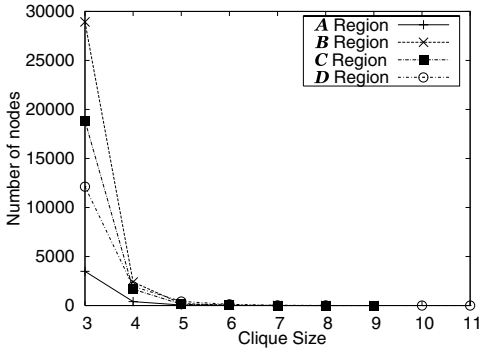


Figure 7: Clique size distribution for all the regions $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$

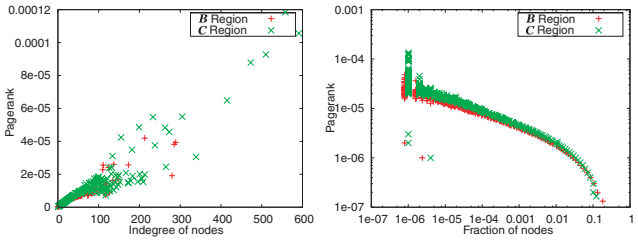
4.6 PageRank

In the context of WWW, the PageRank $p(i)$ of a page i is a measure of citation importance and is defined through the following expression:

$$p(i) = \frac{q}{N} + (1-q) \sum_{j:j \to i} p(j)/d_{out}(j) \qquad i = 1, 2, \ldots, N \quad (1)$$

where N is the total number of nodes, $j \to i$ indicates a hyperlink from j to i, $d_{out}(j)$ is the out-degree of page j and $1 - q$ is the so-called damping factor. The PageRank of a page grows with the in-degree of the page as well as the in-degree of the pages that point to it. It is computed using the algorithm presented in [14]. Similarly, the PageRank value of an individual in a telecom network might indicate the *social importance* of the individual. The social importance of a customer grows with number of people calling that customer as well as the social importance of the callers. For our analysis, we ignore very short calls to remove noise introduced due to wrong numbers and commercial spam.

Figure 8 (a) shows the correlation of the in-degree with the PageRank value of nodes. We observe that PageRank and in-degree are highly correlated. In Figure 8 (b), we observe that PageRank values follow the power law distribution in the network. *Nodes with high PageRank are possibly the ones with high social influence. Since there are only a few of them, the telecom operators can target these influential people to retain them.*



(a) In-degree and PageRank correlation (b) Power law in the distribution of PageRank

Figure 8: Correlation between in-degree and Pagerank for regions $\mathcal{B}$ and $\mathcal{C}$

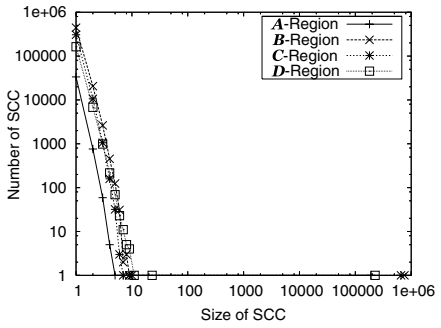


Figure 9: Distribution of SCCs in log-log scale

4.7 Strongly connected components

Scale free graphs usually exhibit the presence of a giant Strongly Connected Component (SCC). With this in mind, we next investigate the distribution of SCCs in the mobile call graphs. Figure 9 shows the distribution of SCC for different regions (in log-log scale). We found that a giant SCC exists in all the call graphs. For example, region $\mathcal{B}$ has an SCC of size 0.7 million nodes. Further, Figure 9 shows that sizes of SCCs closely follow a power law distribution. However, the largest SCC is significantly larger than any of the remaining ones. Consequently, the second largest SCC is very small compared to the largest one. Our results conforms with those obtained for WWW graph [7]. In a later section, we analyze SCC and its association with other components to infer the shape of mobile call graphs.

4.8 Clustering coefficient

Finally, we study the clustering coefficients of nodes in call graphs. Given a triple $< u, v, w >$ of nodes, with mutual relations between v and u as well as between v and w, the clustering coefficient represents the likeliness that u and w are also related, i.e. a friend of a friend is likely to be a friend. Clustering coefficient has important implications in the context of social network analysis [23]. In terms of network topology, it implies the presence of a large number of triangles. The *clustering coefficient* C_i of a node i is defined as:

$$C_i = \frac{\#\text{number of triangles connected to node } i}{\#\text{number of triples centered on node } i}$$

The clustering coefficient of a graph is simply the average clustering coefficient of its nodes. The clustering coefficient

Table 4: Clustering coefficient values

Results from our dataset (directed graphs)				
Region	$\mathcal{A}$-Region	$\mathcal{B}$-Region	$\mathcal{C}$-Region	$\mathcal{D}$-Region
Clus.coeff	0.1461	0.105	0.1667	0.100
Results from other datasets (directed graphs)				
Network	Film-actors	Email	WWW nd.edu	student
Clus.coeff	0.78	0.16	0.29	0.001

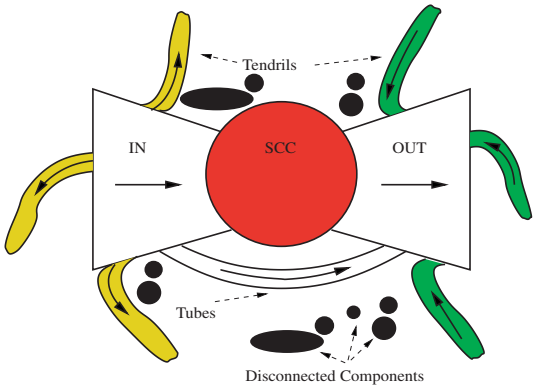


Figure 10: Shape of Bow-Tie Network

for the four regions have been listed in Table 4. We provide comparison with values for other networks [20]. Interestingly, we observe that clustering coefficients of email graphs is similar to those of mobile call graphs. This possibly indicates that the interaction pattern of email as communication mode is similar to the interaction pattern using mobile phones. Further investigation with these triangles showed that there was a large percentage of nodes with $C_i = 1$, which corresponds to cliques of size 3.

5. THE SHAPE OF CALL GRAPHS

Table 5: Random Start BFS Experiment

Reach	Percentage Reach	Reach Probability
< 6	< 0.0005	28.5
1022575	81.7	63
> 1022576	> 81.7	8.5

In this section, we analyze call graphs in order to examine its macroscopic shape. The shape is crucial for two reasons. It provides an intuitive description of the network that is easy to understand and work with, and even more importantly, provides the basis for the development of a generative model. A *generative model*, when found, will provide a valuable simulation tool for Telecom operators to study and predict usage growth in a new region.

In this objective, we do not care whether this structure captures all the characteristic properties, instead it is a way to represent the spatial distribution of edges in the call-graph. Works on the Internet (Jellyfish model [22], Medusa model [8]) and WWW (Bow-tie model [7]) succinctly draw the spatial distribution of nodes. To our knowledge, this is first attempt for revealing the call-graph topology.

We first begin by doing reachability experiments, very similar in spirit to those done for the WWW [7], and supplement those techniques with a few novel ones of our own, in order to expose further details of our call graphs. A crucial insight obtained as a result of this study is that the distribution of *edges* rather than the vertices across the various components leads to a more accurate characterisation of the structure of our mobile call graphs.

We also present ideas on how the knowledge of this structure can be used by Telecom business analysts. While the applicability of the *Treasure-Hunt* model beyond our call graphs is hitherto unknown, all the techniques used in this section are general enough for analyzing any massive graph.

5.1 Structure based on node distribution

Our first goal is to spot all the connected components and place them spatially along with their interconnections to identify the shape. The distribution of strongly connected components is reported in Table 9 for region $\mathcal{B}$ (Figure 9). The results show the existence of one giant strongly connected component.

To discover different components that link *large* connected components, we analyzed our call graph using Random Start Breadth-First-Search (BFS) [11]. Some important definitions are given below.

- **Reach** (R) is the number of all possible nodes reached in BFS, when starting from a given node.
- **Percentage Reach** ($p = R/N$) is the percentage of nodes reached (to total number of nodes in the graph).
- **Reach Probability** (P_R) denotes the percentage probability that a given node has *reach* R.

The experiment collected a set of random sample nodes and computed the *reach* of all these nodes. The various values of the *reach* R is plotted against the number of nodes having *reach* R. The *reach probability* for a given value of *reach* R can be obtained from this distribution.

The experiment conducted on one of the regions ($\mathcal{B}$), produced the results as shown in Table 5. Similar percentages were obtained for the other regions also.

We found that the unique values of *reach* were limited. For instance, for region $\mathcal{B}$, the *reach* was either between 1 to 6 or between 1022575 to 1022586. (For the other regions the reach was similarly split into two ranges). This suggests the existence of a massively connected component CC (nodes having *reach* exactly equal to 1022575), an *entry* component (nodes having *reach* more than 1022575), an *exit* component (nodes have *reach* less than 6) and some disconnected components.

Reachability analysis allows the identification of a strongly connected component SCC) if it exists, and of the regions connected to it. To borrow the terminology of [7], IN refers to the region from which there are paths that leads to the SCC, and OUT refers to the region that is reachable from the SCC. The bow-tie model for the web graph was obtained as a result of reachability analysis, and is named so, because the relative number of vertices in each of the regions IN, OUT and SCC are nearly of the same order reminiscent of a bow-tie.

The bow-tie model(Figure 10), introduced for the WWW, contains a strongly connected component (SCC) region which contains nodes that are mutually reachable, the IN region contains the nodes from which the SCC can be reached. The OUT region contains the nodes that are reachable from the SCC. The $TENDRILS$ gather nodes reachable from

Table 6: Sizes inferred for the bow-tie model

Bow-tie Component	% of total nodes
IN	8.5
SCC	63
OUT	18.7
TENDRIL, TUBE and DISC	9.8

the IN component and reaching neither SCC nor OUT. $TENDRILS$ also include those nodes that reach into the OUT region but do not belong to any of the other defined regions. The $TUBES$ connect the IN and OUT regions directly, and some nodes are totally disconnected ($DISC$).

To find the sizes of SCC, IN and OUT in our call graphs, these components were analyzed for reachability. Starting from any vertex in the IN region, the BFS algorithm reaches all of SCC and OUT region. Hence, all the nodes of IN region have high reachability. From Table 5, we know 8.5 % of nodes reach >1022575 nodes, hence size of IN is 8.5 % of total nodes.

If a starting vertex lies in the SCC region, the BFS cannot reach IN, but can reach the SCC and OUT. Moreover, since all the nodes of SCC are mutually reachable, all vertices in SCC have the same reach. From Table 5, we infer that 63 % of nodes have the same reach of exactly 1022575 (81.7%) nodes. So, the size of SCC is 63% of nodes and the size of SCC and OUT combined should be 81.7 % of nodes, it implies that size of OUT as 18.7 % of nodes.

If the starting vertex happens to fall in OUT, $DISC$, $TENDRIL$ or $TUBE$, then its *reach* should be negligible. Evidently, our experiment showed that around 28.5 % of nodes *reach* 1 to 6 nodes. We summarize our results in Table 6.

We validated our experiments using Pajek [1] tool for this data set. The results matched with the percentages for IN, OUT and SCC region that we obtained. The relative sizes of these regions indicate a structural difference from the sizes of those in the WWW graph. The sizes of IN,SCC,OUT for the WWW are nearly of same order (44 million, 56 million, and 44 million respectively) [7]. For our graphs, the SCC is often an order of magnitude larger than IN, and OUT is often nearly twice that of IN (124801, 755592, 266984 respectively). Hence, the bow-tie model does not characterise our graphs. However, this does not rule out the possibility of another model which is still based on the node distribution.

5.2 Structure based on edge density

To find the shape (hence a model) of our graphs, we examined the the number of vertices in the various regions (IN, OUT,etc.). Though there was some pattern (roughly the same order of magnitude) in the vertex distribution, we found that the corresponding edge distributions among the regions was more striking. We detail this finding now.

From the BFS experiment, we know that starting from a particular node, the reachability is either huge (>1022575) or very low (< 6). We collected the nodes whose reachability is very high. These are the nodes of SCC and IN region. We also collected the nodes that are reached, starting from nodes with high reachability. These nodes belong to the SCC and OUT regions. We intersected these two sets to isolate the SCC, IN and OUT components. With the help of these nodes, we extracted the following edge-induced subgraphs as shown in Table 7.

To understand the structure of call graph, we extracted

Table 7: Definition of subgraphs

Subgraph name	Definition
IN-IN	Subgraph containing the edges only between nodes of IN region
IN-SCC	Subgraph consisting edges one end from IN region and another from SCC
IN-OUT	Subgraph consisting edges one end from IN region and another from OUT
SCC-SCC	Subgraph containing the edges only between nodes of SCC region
SCC-OUT	Subgraph consisting edges one end from SCC region and another from OUT
OUT-OUT	Subgraph containing the edges only between nodes of OUT region

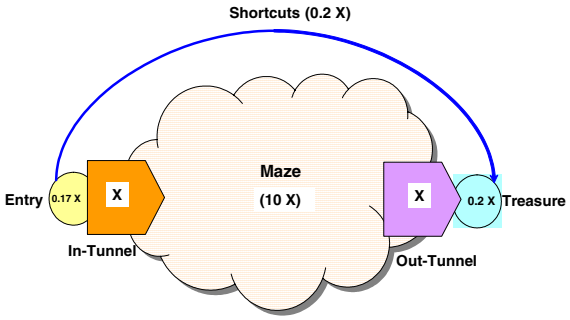


Figure 11: Treasure-Hunt Model

the edge-induced subgraphs of the four regions and studied their properties. Table 8 gives us the results of various parameters that help in detailing the shape of the subgraphs and their boundaries. Most of the columns are self-explanatory. The *Graph type* column gives the kind of subgraph induced from the global graph. For example, edge induced subgraph IN-SCC is a bipartite directed graph as one end is chosen from IN region and another is from SCC region. *Left partition* and *Right partition* capture the number of nodes from the two sets of bipartite graph. The H_{max} *range* column gives a hop range at which maximum neighbours are captured. Entries of the form $H_i = k * H_j$ are provided to give an indication that neighbours at hop i (where the maximum occurs) are k times larger than hop j. The *Edge ratio* column reports the ratio of edges in a particular component to the edges of IN-SCC region. The results for the subgraphs involving disconnected components like IN-DISC, OUT-DISC, DISC-DISC are ignored as the magnitudes of the parameters were negligible.

The ratio of edges in the subgraphs shown in Table 7 display the (cross-region) generic pattern in which these subgraphs connect with each other. The similarity of the edge ratios (see column *Edge ratio* of Table 8) motivated us to present a generic structure capturing the edge ratio of call-graph.

5.3 The *Treasure-Hunt* model

We now introduce and define the *Treasure-Hunt* model which is based on the edge distribution among the various components of a graph and fits our mobile call graphs well. Figure 11 shows the model. Note that X denotes the number of edges in the IN-SCC region.

We chose the treasure-hunt metaphor for describing the model because it captures the shape of the directed graph,

Table 8: Results of Subgraphs of various regions

Reg.	Graph-type	Subgraph	Left part.	Right part.	Edges	Avg d_{in}	Avg d_{out}	H_{max} Range	Dia.	Edge Ratio
$\mathcal{A}$	Directed	IN-IN	4048	-	4525	1.11	1.11	$H_1 = 20 * H_2$	3	0.04X
	Bipartite & dir.	IN-SCC	11043 (IN)	98726 (SCC)	124991	11.31	1.26	1	1	X
	Bipartite & dir.	IN-OUT	1016 (IN)	9943 (OUT)	12154	11.96	1.22	1	1	0.1X
	Directed	SCC-SCC	189327	-	1617431	8.5	8.5	6	10	12.9X
	Bipartite & dir.	SCC-OUT	33855 (SCC)	18873 (OUT)	49649	1.46	2.63	1	1	0.4X
	Directed	OUT-OUT	3396	-	2401	0.71	0.71	$H_1 = 40 * H_2$	2	0.02X
$\mathcal{B}$	Directed	IN-IN	53544	-	55128	1.03	1.03	$H_1 = 10 * H_2$	4	0.17X
	Bipartite & dir.	IN-SCC	110340(IN)	242147(SCC)	311595	2.82	1.28	1	1	X
	Bipartite & dir.	IN-OUT	25948 (IN)	69328 (OUT)	82182	3.17	1.18	1	1	0.26X
	Directed	SCC-SCC	757933	-	3417025	4.5	4.5	7-10	14	10.9X
	Bipartite & dir.	SCC-OUT	291980 (SCC)	239147(OUT)	459724	1.57	1.92	1	1	1.47X
	Directed	OUT-OUT	94287	-	73702	0.78	0.78	$H_1 = 12 * H_2$	4	0.23X
$\mathcal{C}$	Directed	IN-IN	33814	-	36894	1.09	1.09	$H_1 = 12 * H_2$	3	0.11X
	Bipartite & dir.	IN-SCC	77068 (IN)	251170 (SCC)	329651	4.27	1.31	1	1	X
	Bipartite & dir.	IN-OUT	17136 (IN)	52858 (OUT)	64165	3.74	1.21	1	1	0.19X
	Directed	SCC-SCC	658170	-	3351621	5.1	5.1	7-8	12	10.1X
	Bipartite & dir.	SCC-OUT	255050 (SCC)	183309 (OUT)	424299	1.66	2.31	1	1	1.28X
	Directed	OUT-OUT	59013	-	44723	0.76	0.76	$H_1 = 15 * H_2$	4	0.14X
$\mathcal{D}$	Directed	IN-IN	19805	-	18656	0.94	0.94	$H_1 = 10 * H_2$	4	0.17X
	Bipartite & dir.	IN-SCC	52919 (IN)	75441 (SCC)	109711	2.07	1.45	1	1	X
	Bipartite & dir.	IN-OUT	8796 (IN)	15662 (OUT)	19104	2.17	1.22	1	1	0.17X
	Directed	SCC-SCC	226375	-	1123814	4.5	4.5	7-9	13	10.2X
	Bipartite & dir.	SCC-OUT	72858 (SCC)	60763 (OUT)	111626	1.53	1.83	1	1	1.0X
	Directed	OUT-OUT	24355	-	20860	0.85	0.85	$H_1 = 10 * H_2$	4	0.19X

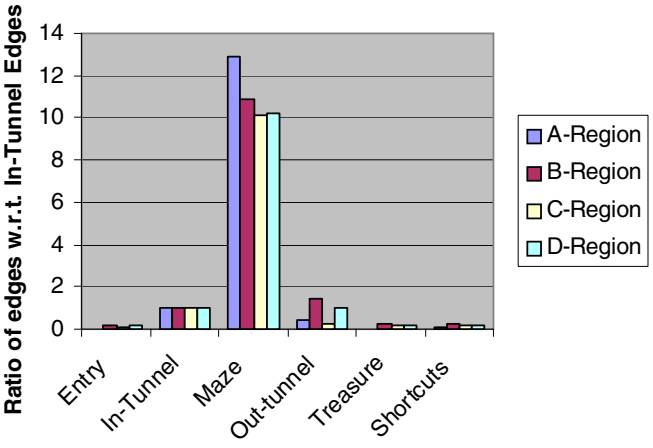


Figure 12: Distribution of edge fractions in different components

and emphasizes the importance of the edges (paths) rather than the nodes. The *entry, in-tunnel, maze, out-tunnel, treasure, shortcuts* are the six regions that are prime components of our *Treasure-Hunt* structure (Figure 11). The region names are also metaphorical and do not imply that the vertices in the *treasure* denote important customers, for example. The smallest of our regions are *entry* and *treasure*. Understandably, not many entries exist, nor there is a lot of treasure. Once, we set a voyage for treasure starting from *entry*, there are lot of obvious paths via *in-tunnel* to the *maze*. The *maze* defines a huge number of convoluted paths, making it harder to reach the treasure. But the chances are fair, there are almost equal number of *out-tunnel* as *in-tunnel* paths to reach the treasure. Interestingly, lucky people may find *shortcuts* that connects the *entry* directly to the *treasure*. (The number of *shortcuts* are as likely as an *entry*). But getting into *entry* and then to *maze* is more than 90% likely.

5.4 Call graph as a Treasure-Hunt structure

The *Treasure-Hunt* model is based on the classification of edges into 6 components. There are some intra-connections (within a component) and interconnections (across components). We *fit* the *Treasure-Hunt* model on the call graphs and tested it with the four regions and found the ratio of edges distributed in all the components (Figure 12). The number of hops within IN-IN and OUT-OUT subgraph shows that there are not many neighbours after 1 hop (see column $H_{max}Range$ in Table 8). So, the IN subgraph can be split into two layers with one of them connecting *SCC*, and another which connects to itself. Obviously, the nodes of IN region splits into two layers as *entry* and *in-tunnel* region in *Treasure-Hunt* model. Similarly, the OUT region is separated into the *out-tunnel* and the *treasure*.

To fit region $\mathcal{B}$ to the *Treasure-Hunt* model, the edge induced subgraphs IN-IN, IN-SCC,SCC-SCC,SCC-OUT,OUT-OUT, IN-OUT can be mapped to *entry*, *in-tunnel*, *maze*, *out-tunnel*,*treasure*, and *shortcuts* respectively. The *edge ratio* column of Table 8 gives the relative magnitude of the edges of each component with respect to the *in-tunnel*. The shape is conclusive as *entry* is 0.16 times *in-tunnel* and *maze* is almost 10 times the *in-tunnel*. The *out-tunnel* is similar in size as *in-tunnel*, whereas *treasure* is relatively the smallest and almost in the same order as *entry*. The *shortcuts* are paths that directly connect *entry* to *treasure*; they are also smaller in magnitude and of the same order as the *entry*.

We tried to fit the other regions $(\mathcal{A}, \mathcal{C}, \mathcal{D})$ (see other edge ratios in Table 8) and found that they fit the *Treasure-Hunt* model quite closely. The *Treasure-Hunt* model brings to light the fact that often the edges rather than the nodes of graphs might follow a pattern, as our call graphs indicate.

There are several implications of the results we obtain through path based model of the call graph. It provides telecom operators with insights on how a certain new service roll-out *might be propagated in the network. For example, the propagation chances would be higher if they target nodes with greater reach (belonging to the* Entry *and* In-Tunnel *re-*

Table 9: Distribution of SCCs

SCC Size	Count
755592	1
9	1
8	3
7	2
6	31
5	124
4	454
3	2629
2	20617
1	443274

gions). Similarly, customers can be segmented based on their placement in the structure.

6. CONCLUSIONS

Over the years, a number of important graph metrics have been proposed to analyze and compare the structure of arbitrary graphs. This paper uses a series of graph structural properties that can be employed in a more systematic approach to dealing with network topologies. We used a carefully chosen set of parameter which reveal mostly connectivity directed characteristics and used them on call graphs. Such metrics can be employed by business strategy planner involved in the telecom domain. We hope that our methods will enable a more rigorous and consistent method of analyzing call graphs and also enable researchers and business community to gain insight into call graphs. These results can significantly affect business strategies.

The shape of the call graph of four disparate regions are in good agreement with the *Treasure-Hunt* model. Although this is promising, only further studies with more call graphs can serve to verify or refute this model.

A problem worthy of consideration is to find a generative model for the mobile call graphs, if one exists. If found, it is likely to offer deep insights into how a mobile operator's customer base evolves with time.

Acknowledgments

The authors would like to thank the anonymous referees for their comments and suggestions.

7. REFERENCES

[1] Pajek: Program for large network analysis. `http://vlado.fmf.uni-lj.si/pub/networks/pajek`.

[2] Abello, J., Pardalos, P., and Resende, M. On maximum clique problems in very large graphs. In *DIMACS Series, 50, American Mathematical Society* (1999), pp. 119–130.

[3] Barabasi, A. Emergence of scaling in complex networks. *Handbook of Graphs and Networks, S. Bornholdt and H. Schuster (Editors)* (2003).

[4] Barabasi, A.-L., and Albert, R. Emergence of scaling in random networks. *Science 286* (October 1999), 509–512.

[5] Bollobás, B., and Riordan, O. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics 1* (2003), 1–35.

[6] Brin, S., and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30*, 1–7 (1998), 107–117.

[7] Broder, A. Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. L. Graph structure in the web. In *Proc. of 9th International Conference (WWW9)/Computer Networks* (2000), vol. 33, pp. 309–320.

[8] Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., and Shir, E. Medusa - new model of internet topology using k-shell decomposition, 2006. http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cond-mat/0601240.

[9] Chan, W.-H. A., and Yao, K. X. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transaction on Evolutionary Computation 7*, 6 (Dec 2003), 532– 545.

[10] Donato, D., Lauraa, L., Leonardi, S., and Millozzi, S. Large scale properties of the webgraph. *The European Physical Journal B 38* (2004), 239–243.

[11] Donato, D., and Leonardi, S. Mining the inner structure of the web graph. In *Eighth International Workshop on the Web and Databases* (2005).

[12] Euler, T. Churn prediction in telecommunications using miningmart. In *Proceedings of the Workshop on Data Mining and Business (DMBiz)* (2005). citeseer.ist.psu.edu/euler05churn.html.

[13] Faloutsos, M., Faloutsos, P., and Faloutsos, C. On power-law relationships of the internet topology. In *SIGCOMM* (1999), pp. 251–262.

[14] Haveliwala, T. H. Efficient computation of pagerank. Technical report, Stanford University, 1999.

[15] Kleinberg, J. Authoritative sources in a hyperlinked environment. In *Journal of ACM* (1999), vol. 46.

[16] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. Structure and evolution of blogspace. *Commun. ACM 47*, 12 (2004), 35–39.

[17] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Trawling the web for emerging cyber-communities. In *WWW '99: Proceeding of the eighth international conference on World Wide Web* (New York, NY, USA, 1999), Elsevier North-Holland, Inc., pp. 1481–1493.

[18] Newman, M. E. J. Assortative mixing in networks. *Physical Review Letters 89* (2002), 208701.

[19] Newman, M. E. J. Mixing patterns in networks. *Physical Review E 67* (2003), 026126.

[20] Newman, M. E. J. The structure and function of complex networks. *SIAM Review 45* (2003), 167.

[21] Palmer, C. R., Gibbons, P. B., and Faloutsos, C. Anf: a fast and scalable tool for data mining in massive graphs. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2002), ACM Press, pp. 81–90.

[22] Siganos, G., Tauro, S. L., and Faloutsos, M. Jellyfish: A conceptual model for the as internet topology. *Journal of Communications and Networks*. Under review.

[23] Watts, D. J., and Strogatz, S. H. Collective dynamics of small-world networks. *Nature 393* (1998), 440–442.

[24] William Aiello, Fan Chung, L. L. A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing* (May 21-23 2000), pp. 171–180.