

# Analyzing the Structure and Evolution of Massive Telecom Graphs

Amit A. Nanavati, Rahul Singh, Dipanjan Chakraborty, Koustuv Dasgupta (*Member, IEEE*), Sougata Mukherjea, Gautam Das, Siva Gurumurthy, Anupam Joshi (*Senior Member, IEEE*)

**Abstract**—With ever growing competition in telecommunications markets, operators have to increasingly rely on business intelligence to offer the right incentives to their customers. Existing approaches for telecom business intelligence have almost solely focused on the individual behavior of customers. In this paper, we use the Call Detail Records of a mobile operator to construct Call graphs, that is, graphs induced by people calling each other. We determine the structural properties of these graphs and also introduce the *Treasure-Hunt* model to describe the shape of mobile call graphs. We also determine how the structure of these call graphs evolve over time. Finally, since Short Messaging Service (SMS) is becoming a preferred mode of communication among many sections of the society we also study the properties of the SMS graph. Our analysis indicates several interesting similarities as well as differences between the SMS graph and the corresponding call graph. We believe that our analysis techniques can allow telecom operators to better understand the social behavior of their customers, and potentially provide major insights for designing effective incentives.

**Index Terms**—Telecom Call Data Records, Graph Algorithms, Social Network Analysis

## I. INTRODUCTION

As Mobile Telecom penetration is increasing, and even approaching saturation in many geographies, 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 [3], [13]. 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: who called whom, when was it made, how long it lasted, 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 cliques or bipartite cores, which implies the presence of communities, can be utilized to improve group targeting and retention.

In this paper we present an analysis of the CDRs of one of the largest Telecom operators in the world. Previously, a few experiments on call graphs of stationary telephone networks had been undertaken to determine parameters like cliques [1] and degree distributions [2]. 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. In particular, we report findings on various topological properties of these massive call graphs, including degree distributions, strongly connected components, and cliques. 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 [21].

One of our primary motivations has been to characterize the shape of call graphs imposed by cellular phone users. For this, we utilize the technique that has been used to arrive at the Bow-Tie model for WWW graphs [6] and then 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 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.

Although several studies on the structure and properties of different types of networks have been reported in the literature most of them are restricted to the analysis of a static snapshot of the network. A temporal analysis of a network to determine how it evolves over time can be very insightful. Such a study is difficult for most real-world networks since it is difficult to obtain information about the arrival of each node and edge into the network. However since the CDRs record the time for all calls, temporal analysis of the Telecom Call graphs is feasible. In this paper we take snapshots of the call graphs at different time segments and analyze how the various parameters of the call graphs change over time.

In several geographies and among many segments of customers, Short Messaging Services (SMS) has become very popular and is a preferred medium of communication. The Call Detail Records maintain the details of all the SMSes and it is possible to create the SMS graph from the CDRs. In this paper we also analyze this graph and try to determine its correlation with the Call graph. Our analysis seems to indicate that the SMS graph is more *social* than the Call graph. As far as we know this is the first study that tries to determine the characteristics of the SMS graph. We believe that insights from the analysis can enable the Telecom operators

understand an important segment of its customers who utilize SMS as a communication medium.

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 present a temporal analysis of mobile call graphs to understand how the various parameters of the graph evolve over time.
- We extend our analysis to SMS graphs and try to correlate the properties between SMS and 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.

The rest of the paper is organized as follows. The next section cites related work. In Section III, we describe our data sets and the techniques for sampling data sources and creation of the call graphs. Various characteristics providing critical insights in the call graphs are presented in Section IV. We then introduce a framework to get the finer level details of the topology in Section V. In Section VI we present a temporal analysis of the call graphs and Section VII extends the analysis to SMS graphs. Finally, we conclude the paper with Section VIII.

## II. BACKGROUND AND RELATED WORK

Massive graphs originating from different sources like WWW, Internet topology, Email graphs and Biological networks have drawn the attention of a plethora of researchers [21], [12], [7]. 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 which characterize the graphs based on various parameters like size, density, degree distribution, clustering coefficient and connected components, 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. [14] showed that the degree distribution of the internet follow a Power law. Both [4] and [18] suggest that the *in* and *out* degrees of vertices on the Web graph also 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. On-line friendship networks also exhibit the small-world phenomenon [20]. Our analysis reveals evidence of small world phenomenon in mobile graphs also.

Determining groups of related pages in such networks is another interesting problem. For example, [18] showed that *bipartite cores* in the Web graph represent implicitly defined communities. 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* [5] 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 [15] who defined two types of scores for

Web pages which pertain to a certain topic: *authority* and *hub* scores. We also try to identify communities and important mobile numbers from the Telecom graphs.

Yet another body of work has been undertaken to determine topological model of the WWW graph. [6] 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. The Daisy model [11] is an attempt to further refine the WWW bow-tie model. Researchers have also tried to determine models for the Internet topology. The Jellyfish model [24] was one of the first in this direction. The Medusa model [8] is yet another model for the Internet topology which was derived using a technique called *k*-core decomposition. In this paper we introduce a model for Telecom graphs.

Although real-world graphs are evolving over time most of the analysis reported in the literature have been done on static graphs. Recently, structural properties of different snapshots of the WWW graph has been reported in [22]. [19] showed that Citation graphs exhibit densification and shrinking diameters over time and presented a generative model called the *Forest Fire model* for capturing this phenomenon. On the other hand, the emergence of bursty communities in the blogspace was identified in [16]. The structure and evolution of two Online Social Network maintained by Yahoo has been reported in [17].

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 [1]. 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. [2] 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  $\alpha$ - $\beta$  graphs to capture the distribution. In this paper we present deeper insights on the characteristics of Telecom call and SMS graphs to compliment the earlier studies.

In terms of business strategy design for telecommunication industry, many existing techniques exist based on mining of user profiles [13] as well as application of machine learning methods [3]. 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 prevention.

## III. 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 and generally stored in a Data Warehouse. In this paper we analyze the CDRs of one of the largest mobile operator in India. Not surprisingly, for a large country like India with a significant mobile penetration, more than a billion calls are made every month, and the data storage runs into terabytes.

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

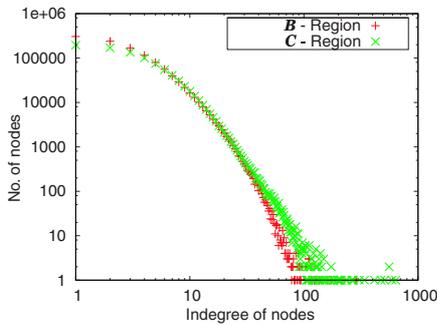


Fig. 1. In-degree distribution for regions  $\mathcal{B}$  and  $\mathcal{C}$  ( $\gamma_{in}$  is 2.85924 and 2.89961)

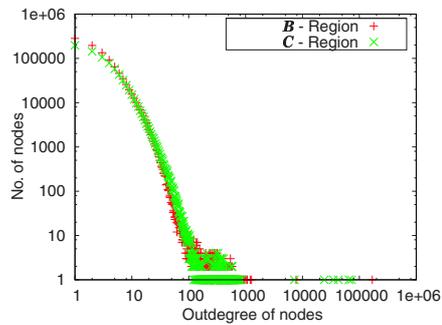


Fig. 2. Out-degree distribution for regions  $\mathcal{B}$  and  $\mathcal{C}$  ( $\gamma_{out}$  is 1.71374 and 1.70808)

TABLE I  
DETAILS OF DATA SET USED

Region	Nodes	Edges	Period	Avg. Deg.	Type
$\mathcal{A}$ -Region	224418	1816285	1 month	16.19	Directed
$\mathcal{B}$ -Region	1250656	4514528	1 week	7.22	Directed
$\mathcal{C}$ -Region	989573	4313797	1 week	8.72	Directed
$\mathcal{D}$ -Region	407332	1456645	1 month	7.15	Directed

of  $G$ , then an edge  $\langle u, v \rangle$  is said to exist if  $u$  calls  $v$ . Such a graph can be easily constructed from the CDRs.

To get the actual graph, we had to query the Data warehouse 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 in India.
- The mobile operator has broken the country into several region-wise circles. For our study, we analyzed the calling patterns of four circles; two of these circles were large metropolitan cities and two were states with a mixture of rural and urban population. These regions are in different areas of India and are also culturally diverse with different local languages.
- 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 the 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. (However in many countries some of these missed calls may be used to convey pre-determined messages and are therefore socially relevant).
- 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.

Table I 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  $\bar{k} = 2m/n$ .

## IV. 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 [10], [9], [23] for computing these parameters.

### A. 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 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  $\gamma$  is between 2 and 3, very much like the WWW graph [6]. 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.*

### B. Neighbourhood Distribution

The neighbourhood function,  $N(h)$  for a graph also called hop-plot [14], 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

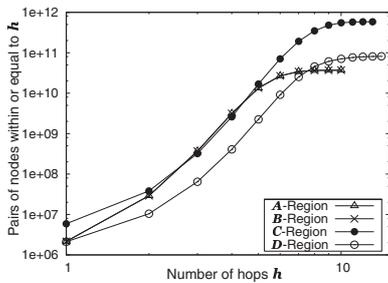


Fig. 3. Neighbourhood function distributions for all regions

TABLE II

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	$\mathcal{H}$	$C$	$\delta_{\text{eff}}$
$\mathcal{A}$	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

follows:

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

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 neighbourhood function provides us ways to compare call graphs in terms of *hop-exponent*, distance distribution, and effective diameter.

We calculated the neighbourhood function for call graphs using the ANF tool [23]. The plot of the neighbourhood function for all hop for all the regions are shown in the Figure 3. Also,  $N(h) \propto h^{\mathcal{H}}$ , where  $\mathcal{H}$  is the *hop 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 [23].

We computed the *hop-exponent* using linear fit on the neighbourhood function distribution shown in Figure 3. The hop exponents of the 4 regions are reported in the Table II. We consistently found hop exponent close to 4 and 5 in the Telecom graphs. 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 [14]. *This suggests that our mobile telecom graphs are structurally as dense as those of the Internet graphs. Interestingly, even though the graph of regions A and B differ considerably in parameters such as average degree, number of nodes and edges (Table I), their hop exponents are similar (3).*

**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 [14]:

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

The effective diameter of a network is  $\delta_{\text{eff}}$  if any two nodes are within  $\delta_{\text{eff}}$  hops from each other *with a high probability*. The effective diameter for all the four regions are given in the Table II; the maximum value is 13. This provides evidence of small-world phenomenon in mobile call graphs since most pairs of nodes (phone numbers) are separated by a handful of edges (calls). We believe that this phenomenon can be further exploited to identify (social) communities.

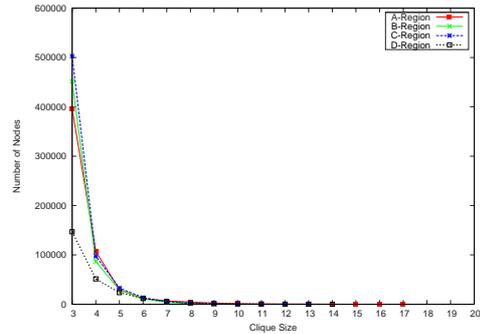


Fig. 4. Clique size distribution for all regions

### C. 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 for the 4 regions under observation are shown in Figure 4. Cliques of size as high 17 is observed. Relatively, large number of cliques of smaller sizes like 3 and 4 are also observed.

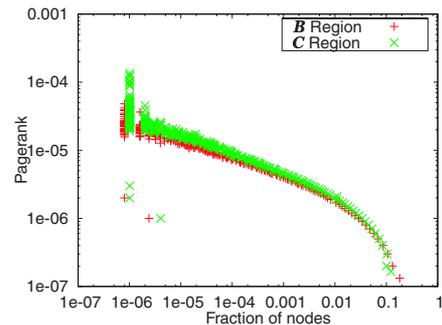


Fig. 5. Power law in the distribution of PageRank

### D. PageRank

In the context of WWW, the PageRank  $p(i)$  [5] of a page  $i$  is a measure of citation importance and is defined through the following expression:

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

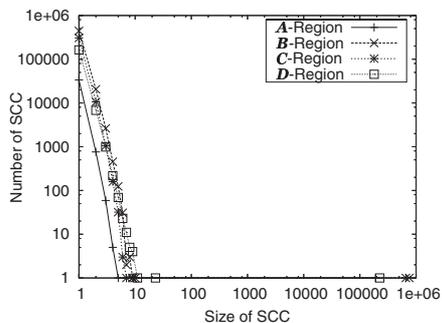


Fig. 6. Distribution of SCCs in log-log scale for all regions

where  $N$  is the total number of nodes,  $j \rightarrow i$  indicates a hyperlink from  $j$  to  $i$ ,  $d_{out}(j)$  is the out-degree of page  $j$  and  $q$  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.

Figure 5 shows the PageRank distribution for region  $B$  and  $C$  respectively. For our analysis, we ignore very short calls to remove noise introduced due to wrong numbers and commercial spam. We observe that PageRank values follow the power law distribution in the network.

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. *Since there are only a few of nodes with high PageRank, the telecom operators can target these influential people to retain them.*

### E. 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 6 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  $B$  has an SCC of size 0.7 million nodes. Further, Figure 6 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 conform with those obtained for WWW graph [6]. In the next section, we analyze SCC and its association with other components to infer the shape of mobile call graphs.

## V. THE SHAPE OF CALL GRAPHS

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. To our knowledge, this is the first attempt for revealing the call-graph topology.

We first begin by doing *reach* experiments, very similar in spirit to those done for the WWW [6], and supplement those techniques with a few novel ones of our own, in order to expose

TABLE III  
Distribution of SCCs for region  $B$

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

further details of our call graphs. Based on our analysis we present a model called *Treasure-Hunt* for mobile 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 characterization of the structure of the 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.

### A. 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 III for region  $B$ . 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]. While ‘reachability’ of a node  $v$  means whether  $v$  is reachable from another node  $u$ , we use ‘reach’ of  $v$  to mean the set of nodes (or its cardinality) reachable from  $v$ . Some important definitions are given below.

- **Reach** ( $R$  of a node  $u$ ) 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*  $R$ .

TABLE IV  
RANDOM START BFS EXPERIMENT FOR REGION  $B$

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

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 ( $B$ ), produced the results as shown in Table IV. Similar percentages were obtained for the other regions also.

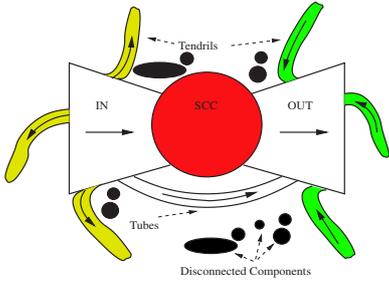


Fig. 7. Shape of Bow-Tie Network

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.

Reach 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 [6],  $IN$  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 reach 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 7), 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 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 reach. 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 reach. From Table IV, 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  $OUT$  regions. Moreover, since all the nodes of  $SCC$  are mutually reachable, all vertices in  $SCC$  have the same reach. From Table IV, 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 since 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 V.

We validated our experiments using the Pajek [9] tool for this data set. The results matched with the percentages for  $IN$ ,  $OUT$

TABLE V  
SIZES INFERRED FOR THE BOW-TIE MODEL FOR REGION  $\mathcal{B}$

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

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) [6]. 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, and perhaps not surprisingly, the bow-tie model does not characterise our graphs. However, this does not rule out the possibility of another model which is based on the node distribution.

TABLE VI  
DEFINITION AND TYPES FOR SUBGRAPHS

Subgraph	Definition	Graph Type
IN-IN	Subgraph containing edges only between nodes of IN region	Directed
IN-SCC	Subgraph consisting edges from IN region to SCC	Bi-partite & Directed
IN-OUT	Subgraph consisting edges from IN region to OUT	Bi-partite & Directed
SCC-SCC	Subgraph containing edges only between nodes of SCC region	Directed
SCC-OUT	Subgraph consisting edges from SCC region to OUT	Bi-partite & Directed
OUT-OUT	Subgraph containing edges only between nodes of OUT region	Directed

## B. 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 reach is either huge ( $>1022575$ ) or very low ( $< 6$ ). We collected the nodes whose reach is very high. These are the nodes of  $SCC$  and  $IN$  region. Starting from nodes with high reach, we collected the nodes that are reachable. 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 several edge-induced subgraphs which are defined in Table VI. 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.

To understand the structure of call graph, we extracted the edge-induced subgraphs of the four regions and studied their properties. Table VII 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. *Left partition* and *Right partition* capture the number of nodes from the two sets of

TABLE VII  
PROPERTIES OF SUBGRAPHS OF VARIOUS REGIONS

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

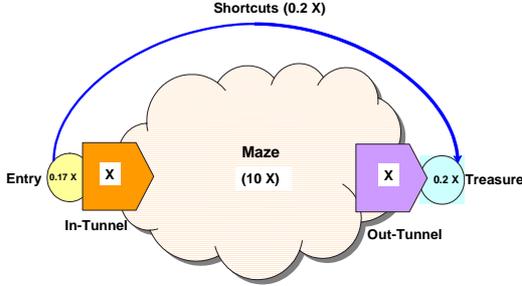


Fig. 8. Treasure-Hunt Model

bipartite graph. 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 VI 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 VII) motivated us to present a generic structure capturing the edge ratio of call-graph.

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 8 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, 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 8). 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

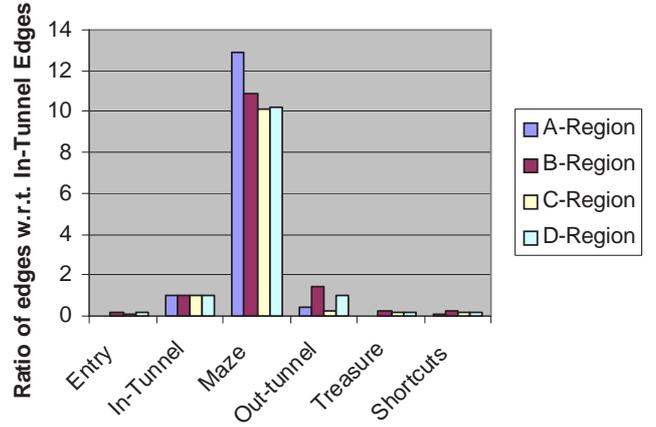


Fig. 9. Distribution of edge fractions in different components

lot of treasure. Once, we set on 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.

### C. 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 of the four regions and found the ratio of edges distributed in all the components (Figure 9). Determining the number of hops within IN-IN and OUT-OUT subgraph shows that there are not many neighbours after 1 hop. So, the IN subgraph can be split into two layers with

one of them connecting *SCC*, and another which connects to itself. Thus, the nodes of IN region split 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 and IN-OUT can be mapped to *entry*, *in-tunnel*, *maze*, *out-tunnel*, *treasure*, and *shortcuts* respectively. The *edge ratio* column of Table VII 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 VII) 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 regions). Similarly, customers can be segmented based on their placement in the structure.*

## VI. TEMPORAL ANALYSIS

In this section we discuss temporal analysis of the call graphs. We studied how some of the structural properties of these call graphs vary with time. For regions  $\mathcal{B}$  and  $\mathcal{C}$  for which we had one week's data we looked at the cumulative call data records at each of the seven days. For regions  $\mathcal{A}$  and  $\mathcal{D}$  for which we had one month's data we looked at seven time points at intervals of four days each. Note that the analysis was done using the CDRs from a period in which there was no celebrations or special events to ensure that the snapshot reflected the normal calling patterns of the customers.

### A. Degree Distributions

The changes in the degree distribution of the graphs over time may yield some insights into the evolution of call graphs. For instance, it might be interesting to find out how the number of nodes with high in-degree and high out-degree changes with time.

The plots for the temporal variation in-degree distributions for regions  $\mathcal{B}$  and  $\mathcal{C}$  on a log-log scale are shown in Figure 10(a) and Figure 10(b) respectively. Regions  $\mathcal{A}$  and  $\mathcal{D}$  showed the same trends. From the results it is evident that the indegrees of all nodes increase with time. The plots for the out-degree distributions for region  $\mathcal{B}$  and  $\mathcal{C}$  are shown in Figure 11(a) and Figure 11(b) respectively. Regions  $\mathcal{A}$  and  $\mathcal{D}$  showed the same trends. The temporal variation of the out-degree distribution is similar to the in-degree distribution.

A network is said to exhibit the *Preferential attachment* [4], [21] property if in the network nodes with higher degree have stronger ability to grab new links. We were interested in finding out if preferential attachment is coming into play in call graphs.

TABLE VIII  
TEMPORAL VARIATION OF CLIQUES OF VARIOUS SIZES FOR REGION  $\mathcal{B}$

Size	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
3	53966	123798	192187	261006	329891	396233	451350
4	5499	18487	33098	48281	64020	77758	86651
5	659	3823	8357	14392	20869	26469	28844
6	54	660	2174	4341	7094	10074	11037
7	6	121	493	1172	2179	3350	3803
8	0	17	88	285	645	953	1236
9	0	0	17	54	197	382	427
10	0	0	0	11	37	104	159
11	0	0	0	0	4	16	41
12	0	0	0	0	2	4	4

For this, we first found out the in-degrees and out-degrees of the nodes on the first day and for the same set of nodes, found the average of their in-degrees and out-degrees on the seventh day. Figures 12(a) and 12(b) plot the in-degree statistics for regions  $\mathcal{B}$  and  $\mathcal{C}$  nodes on the log-log scale. Figures 13(a) and 13(b) plot the out-degree statistics for the same regions. Remarkably the plots show a linear trend on the log-log plot indicating that nodes with higher degrees on the first day also have higher degrees on the seventh day. This gives a clear evidence of preferential attachment being exhibited.

### B. Neighborhood distribution

Next, we analyzed the neighborhood function of the call graphs over time. Since the neighborhood function is an indication of the effective diameter of a graph, this plot gives insights on how the diameter of the call graphs is changing with time.

Figure 14(a) and Figure 14(b) show the neighborhood functions at seven different points in time for the regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively. It is clear from the results that the maximum distance between any two pairs in the graph is decreasing with time. This phenomenon of decreasing diameters has been observed in other graphs as well [19].

### C. Cliques

Table VIII shows the cliques of various sizes that are present in the call graph in each of the seven days for region  $\mathcal{B}$ . By the first day itself the largest sized clique is 7. By the fifth day a clique of size 12 is formed. However no cliques of size greater than 12 are formed in the last two days. The table also shows that there is an almost linear increase in the number of cliques of smaller sizes each day.

### D. Strongly Connected Components

We have also looked at the evolution of the number and sizes of the strongly connected components present in these graphs. Figure 15(a) and Figure 15(b) show the plots for the fraction of nodes present in strongly connected components of size 1,2,3,4,5 and the strongly connected component of the largest size, for the regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively.

It was observed that the fraction of nodes present in the largest strongly connected component increases rapidly with time. Moreover, the fraction of nodes present in strongly connected components of the smallest sizes is decreasing with time.

In [17] similar studies were conducted on the fraction of nodes present in connected components of various sizes in the graph

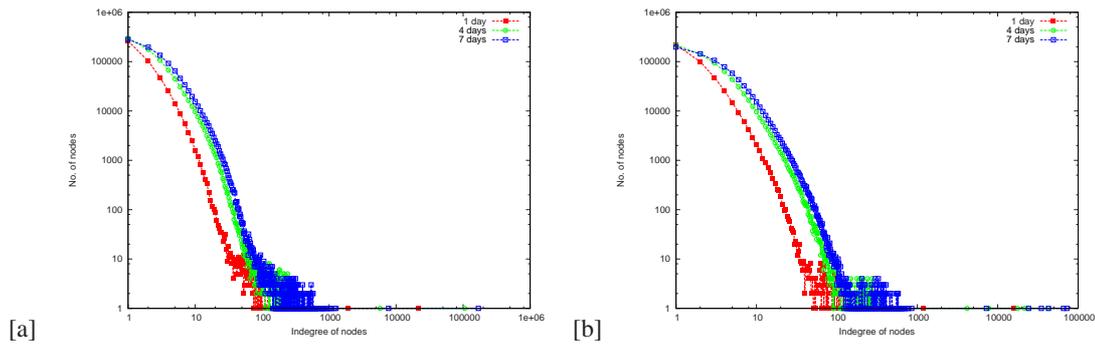


Fig. 10. Temporal variation of in-degree distributions for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

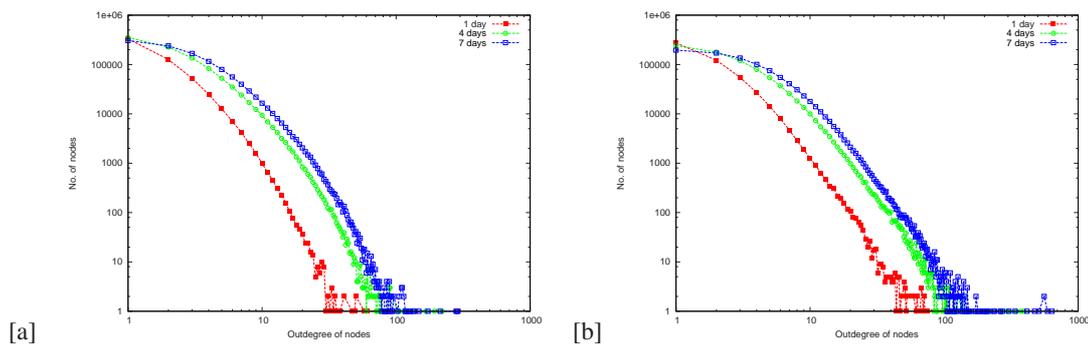


Fig. 11. Temporal variation of out-degree distributions for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

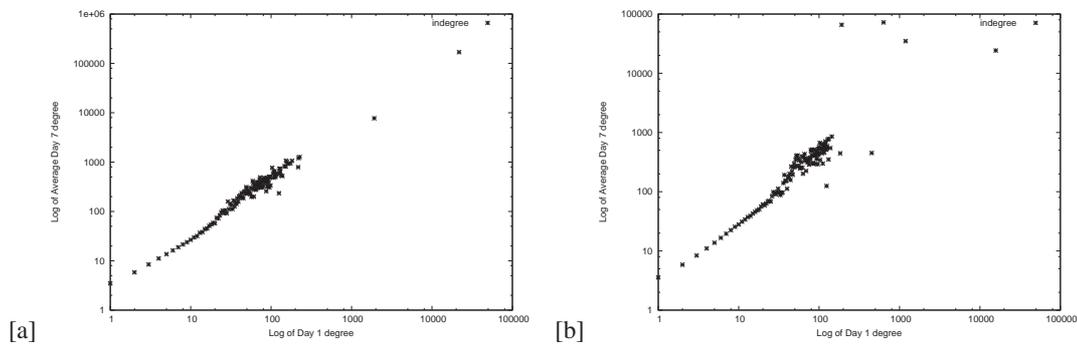


Fig. 12. Evidence of preferential attachment of indegree for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

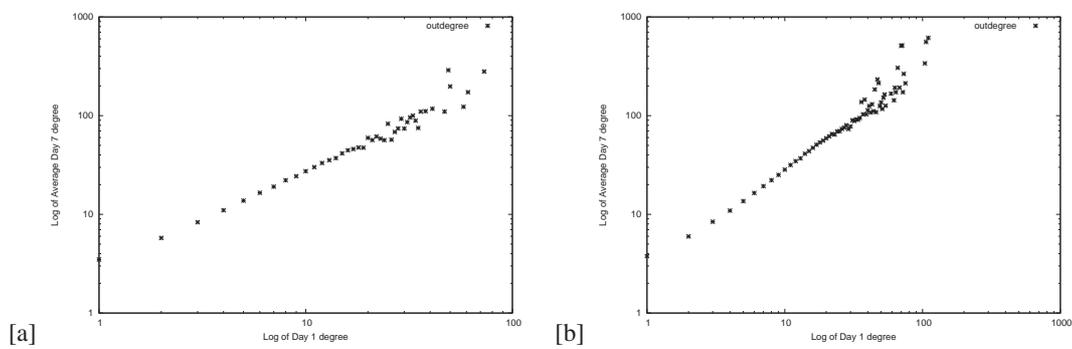


Fig. 13. Evidence of preferential attachment of outdegree for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

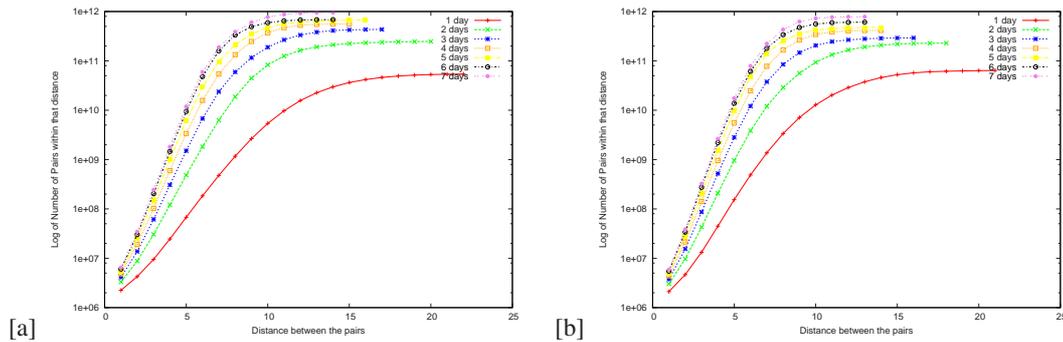


Fig. 14. Temporal variation of neighborhood function distributions for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

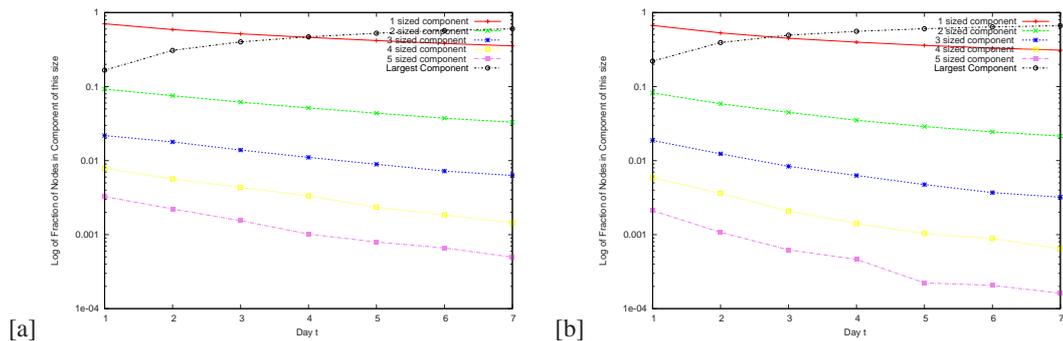


Fig. 15. Temporal variation of the fraction of nodes in strongly connected components of various sizes for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

structure of social networking sites Flickr and Yahoo! 360. An analysis of these graphs revealed that the fraction of nodes in all components was continuously increasing. However, call graphs seem to show a tendency of greater accumulation into a single strongly connected component over time by taking in nodes from the smaller components.

### E. The Treasure-Hunt model

It would be interesting to determine how the *Treasure-Hunt* model that we proposed earlier for call graphs evolves over time. Therefore we fitted the *Treasure-Hunt* model on the graph as it stood at various points of time. Figure 16(a) and Figure 16(b) show the edges present in the various components of the *Treasure-Hunt* model at seven points of time for the regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively. It is interesting to note the striking similarity in the two plots in spite of them belonging to different regions. Some of the trends that are evident from these plots are:

- The number of edges in the maze increase very rapidly. This observation should be expected since we had earlier observed that the percentage of nodes in the strongly connected component was increasing.
- The sizes of the in-tunnel and out-tunnel are also increasing. However the increase is not as rapid as in the case of the maze.
- The sizes of the treasure and entry components are decreasing. Hence the increase in maze is coming at the expense of the treasure and entry becoming smaller. The maze is getting bulkier by sucking in edges from the side components i.e. the entry on the one end and the treasure on the other end.

TABLE IX  
DISTRIBUTION OF NEW NODES TO VARIOUS COMPONENTS OF THE BOW-TIE MODEL FOR REGION  $\mathcal{B}$

Day Number	Fraction of new nodes in IN	Fraction of new nodes in SCC	Fraction of new nodes in OUT
2	0.300206	0.0777751	0.150014
3	0.324856	0.07414	0.173376
4	0.334607	0.0742626	0.184446
5	0.361123	0.0812633	0.213224
6	0.36397	0.0869003	0.22547
7	0.346742	0.0855269	0.233377

- The number of shortcuts remains almost constant with time.
- While studying the *Treasure-Hunt* model we had noticed that the ratio of edges in various components with respect to the edges in the in-tunnel component were surprisingly similar for different regions. We wanted to determine how this ratio changes with time. Figure 17(a) and Figure 17(b) show the values of the ratio of edges in various components to the edges in the in-tunnel component at seven points in time. As expected these plots indicate that:
- The fraction of edges in the maze is increasing.
  - The fraction of edges in entry and treasure are decreasing
  - The fraction of edges in shortcuts and out-tunnel remain almost constant.

1) *Densification in the Treasure-Hunt model*: Figures 16 and 17 show that the maze, entry and treasure component sizes are flattening. We also observed that the size of the maze is continuously increasing while the size of the entry and treasure

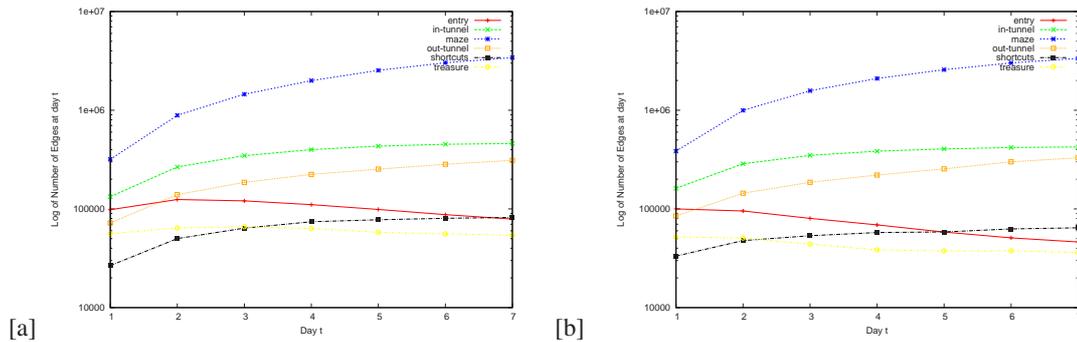


Fig. 16. Temporal variation of the edges in the various *Treasure-Hunt* components for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

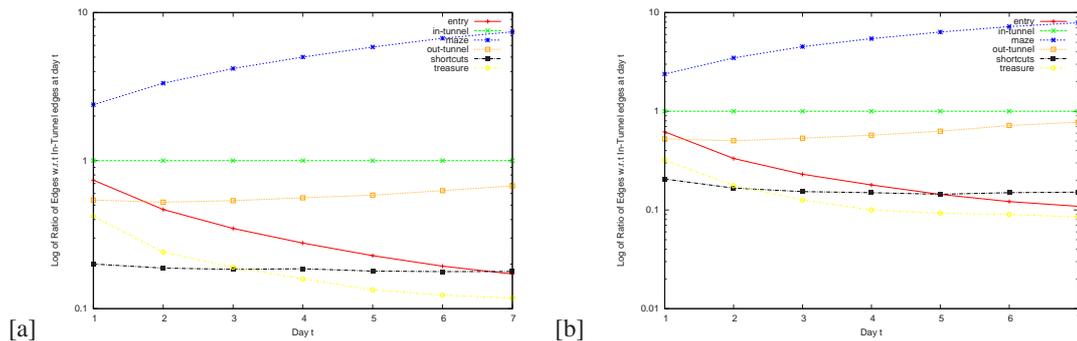


Fig. 17. Temporal variation of the edge ratios of the different components of the *Treasure-Hunt* w.r.t. the *In-tunnel* for regions  $\mathcal{B}$  and  $\mathcal{C}$  respectively

is continuously decreasing. A natural question to ask is whether the graphs shall continue to preserve the ratios of the *Treasure-Hunt* model or will the continued densification lead the various components to collapse into a single large maze.

Each day of the analysis new nodes are added to the Call graph. Table IX shows the fraction of the new nodes that went into the three bow-tie regions IN, SCC and OUT for region  $\mathcal{B}$ . It is clear from the table that a large portion of these new nodes are going into the IN and OUT regions. This indicates that the tremendous increase in the maze is primarily because of older nodes from the other regions being pulled into the maze and not due to the addition of new nodes. The pattern of new nodes coming into the IN and SCC regions also validates the observed calling behavior in the society. The new people who join the network initially make or receive a few calls and hence are part of the IN or the OUT region. Over time they make and receive more calls thus pulling them into the SCC. The constantly high influx of new nodes into the IN and the OUT regions suggests against the total vanishing of the treasure and entry regions.

Another important point that should be noted is that in the current study, due to the shorter interval of analysis, we do not consider the deletion of old edges. Over a period of very long intervals, the call graph accumulates a number of edges which have become stale (no call has been made between the corresponding nodes for a long time). Filtering the call graph to select only edges which correspond to sufficiently frequent and recent calls before analyzing the call graph may give better insights.

If the densification disturbs the ratios in the *Treasure-Hunt*

TABLE X  
DETAILS OF DATA SET USED FOR CALL AND SMS GRAPH COMPARISON.

Region	Nodes	Edges	Period	Avg. Deg.	Type
B-Region Call Graph	2324556	6875678	1 week	2.96	Dir.
B-Region SMS Graph	1162457	2237482	1 week	1.92	Dir.

model, how can we explain the findings of Section V-B which showed that the various regions which were analysed over different periods of time have the same ratios? One plausible explanation is the following: One common feature in these graphs is the order of the number of edges. Since the *Treasure-Hunt* model is an edge based model, it is an indication of the fact that the ratios are a function of the number of edges in the graph.

## VII. SMS GRAPH ANALYSIS

In the previous sections we have concentrated on analyzing the graph induced by voice calls made between mobiles within a particular region. Short Messaging Service (SMS) is very popular in certain geographies and among some customer segments. The Call Detail Records contain information of all the SMSes that are sent and the SMS graph can be extracted from the CDRs. In this section we analyze SMS graph and compare and contrast the structural properties of SMS and Call graphs.

### A. Data Sources

We studied the structural properties of the Call and the SMS graph for the same region over the same period of time. We

selected region  $B$  for this set of experiments. Table X shows the details for the data set used for the experiments which will be reported hereafter.

*It is interesting to note that for the same period of time the number of edges in the SMS graph are only one-third of the number of edges present in the call graph. Similarly the number of nodes is only half of those in the call graph. This indicates that a large number of mobile phone users only make voice calls.*

### B. Structural Properties

In this section we will analyze the SMS graph for some insightful structural properties. In order to compare how these properties differ with respect to the call graphs we will also study the same properties for the call graph.

1) *Reciprocity*: The reciprocity of a directed graph is defined as the fraction of the total edges such that if  $(a, b)$  is an edge in the graph then  $(b, a)$  is also an edge in the graph. From the perspective of Telecom graphs, pairs of nodes which have edges in both directions may be an indication of greater social ties between those nodes as compared to pairs which have only an one-way edge. We found out the reciprocity of the call graph to be 0.3 while that of the SMS graph was almost double with a value of 0.6. *This is an important observation and indicates that there is a greater fraction of social edges in SMS graph as compared to call graphs.* It is interesting to note that the reciprocity of two Social Networking Web sites were found to be around 0.7 and 0.84 [17].

2) *Degree Distributions*: Figure 18(a) and Figure 18(b) show the in-degree distribution for the call and SMS graphs while Figure 19(a) and Figure 19(b) show the out-degree distribution for the graphs

The power-law behavior is apparent for the Call graphs. However for the SMS graph the in-degree and out-degree distributions do not show a marked power-law behavior. The SMS in-degree and out-degree distributions consist of two different regions of linear growth. A sudden kink where the two linear regions meet is evident in both the plots. Thus the degree distributions of SMS graphs are different from many networks that has been studied in the literature.

3) *Neighborhood Distribution*: Figure 20(a) and Figure 20(b) show the neighbourhood function distributions for the call and SMS graphs. The neighborhood function plots for both the graphs appear similar. This indicates that the diameters of both the call graph and the SMS graph are almost the same.

4) *Cliques*: Table XI and Table XII give the number of cliques present in the call graph and SMS graph respectively. The number of smaller size cliques is higher in the Call graph. This is to be expected since the Call graph has a higher number of nodes and edges. However it is very interesting to note the number of higher sized cliques (of size 6,7,8) are higher in the SMS graph in spite of that graph having only one-third the number of edges in the call graph.

*This observation leads to the inference that there are some cliques where the participating members only send SMSes to one another and do not make calls. Such kind of a behavioral pattern may be indicative of a group of people with common interests where the members send bulk SMSes (jokes, etc.) to all the other members of the clique for instance. This statistics indicates that Telecom Operators need to analyze the SMS graphs*

*also to identify an important segment of customer communities who may not be discovered by Call graph analysis.*

TABLE XVI  
RATIO OF EDGES OF CALL AND SMS GRAPH FOR THE DIFFERENT  
TREASURE HUNT SUBGRAPHS

Subgraph	Graph type	Call Graph Edges	SMS Graph Edges	Ratio
IN-IN	Directed	113541	35832	3.17
IN-SCC	Bi-partite & Directed	877417	110280	7.96
IN-OUT	Bi-partite & Directed	203335	12879	15.79
SCC-SCC	Directed	4444481	1466544	3.03
SCC-OUT	Bi-partite & Directed	861238	157932	5.45
OUT-OUT	Directed	138177	73619	1.88
DISC	Directed	237489	380396	0.62
TOTAL	Directed	6875678	2237482	3.07

5) *Strongly Connected Components*: Table XIII and Table XIV give the number of strongly connected components present in the SMS graph and the call graph respectively. The tables show the number of strongly connected components present of each size and also the fraction of the total nodes that are present in the strongly connected components of that size.

For the call graph the largest strongly connected component comprises more than of half the nodes. Apart from this largest strongly connected component, the call graph has strongly connected components of only 10 other sizes and the size of the second largest strongly connected component is only 11. On the other hand, Table XIII reveals that there are a large number of nodes which are present in the smaller strongly connected components in the case of the SMS graph. This graph also has one very large strongly connected component which contains about 37% of the nodes.

*This analysis gives an indication of the fact that the SMS graph is composed of a number of small islands which are strongly connected. This may be due to the fact that during SMSing the people are often part of a community which has lesser links outside.*

### C. The Treasure-Hunt Model

We tried to fit the *Treasure-Hunt* model on the SMS graph. Table XV gives the details of the various components of the model found in the call graph and the SMS graph. On the other hand Table XVI gives the ratio of edges in the call graph compared to the SMS graph for the various subgraphs of the Treasure Hunt.

The call graph has three times the number of edges than the SMS graph. The ratio of edges for the *IN-IN* and *SCC-SCC* subgraphs of the call graph to the SMS graph is also three times. On the other hand the ratios of the edges in the subgraphs *IN-SCC*, *IN-OUT* and *SCC-OUT* of the call graph to the SMS graph are higher. Note that these subgraphs are bipartite and thus have edges in one direction only. Since the SMS graph has higher reciprocity, the proportion of such edges are lower. On the other hand the SMS graph has more edges in *DISC*, the disconnected components. This is also to be expected since we have observed that the SMS graph has many disconnected islands. In fact, despite having fewer edges, it has a higher number of cliques, SCCs, disconnected components, and smaller sized bipartite components, which indicate more reciprocity. **Therefore, the SMS graph follows the Treasure-Hunt Model but with a higher reciprocity.**

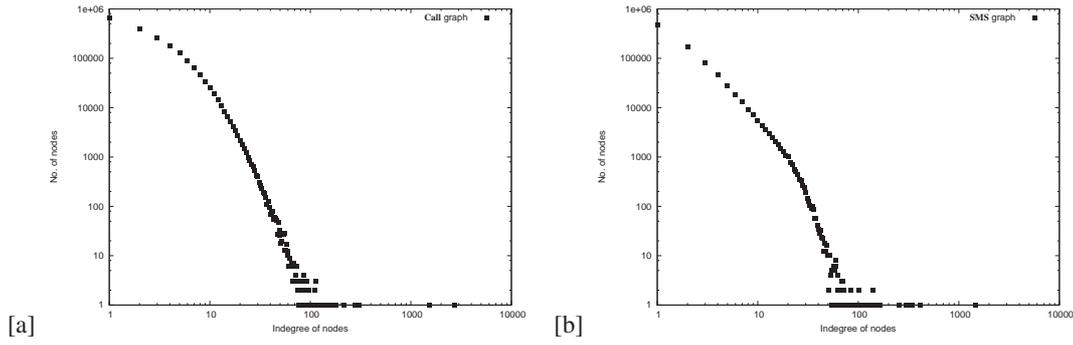


Fig. 18. In-degree distributions for Call and SMS graphs respectively

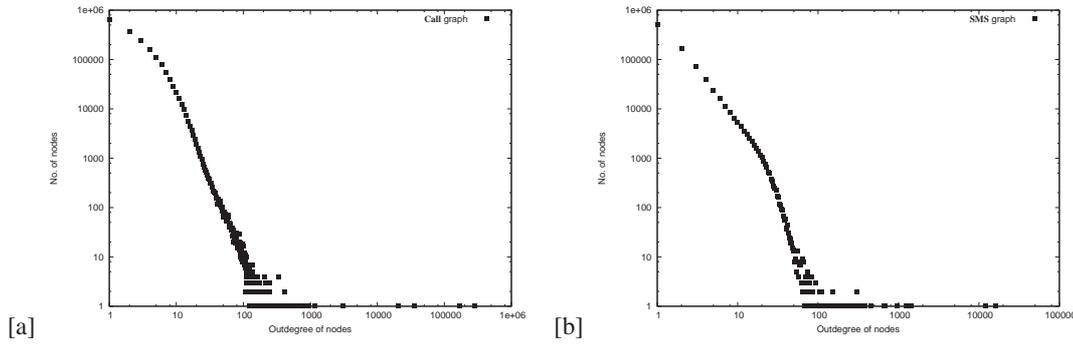


Fig. 19. Out-degree distributions for Call and SMS graphs respectively

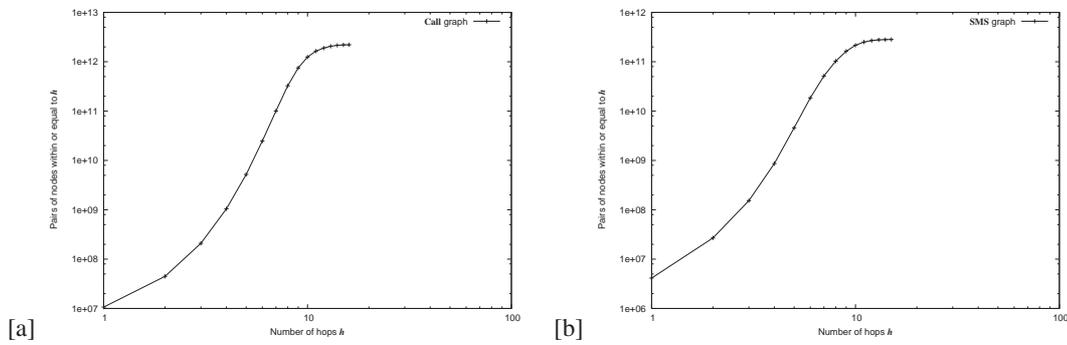


Fig. 20. Neighborhood function distributions for Call and SMS graphs respectively

Clique Size	Count
3	541771
4	32248
5	2543
6	291
7	23
8	17
9	17
10	6
11	2

TABLE XI  
Distribution of cliques for Call graph

Clique Size	Count
3	99910
4	13250
5	2307
6	502
7	129
8	30
9	13
10	5

TABLE XII  
Distribution of cliques for SMS graph

Size	Count	Node Fraction
439444	1	0.37803
22	2	3.78509e-05
21	1	1.80652e-05
20	1	1.72049e-05
18	1	1.54844e-05
17	3	4.38726e-05
16	2	2.75279e-05
15	5	0.000178931
12	17	0.00017549
11	40	0.000378509
10	45	0.000387111
9	92	0.000712284
8	146	0.00100477
7	259	0.00155963
6	520	0.00268397
5	1077	0.00463243
4	2719	0.00935604
3	8580	0.0221428
2	40828	0.0702443
1	590696	0.508144

TABLE XIII  
Distribution of SCCs for SMS graph

Size	Count	Nodes Fraction
1235077	1	0.531317
11	1	4.73209e-06
10	1	4.3019e-06
8	6	2.06491e-05
7	11	3.31246e-05
6	79	0.00020391
5	217	0.000466756
4	919	0.00158138
3	4737	0.00611343
2	33224	0.0285852
1	1003439	0.431669

TABLE XIV  
Distribution of SCCs for Call graph

TABLE XV  
PROPERTIES OF TREASURE-HUNT SUBGRAPHS OF THE CALL AND THE SMS GRAPHS

Type	Subgraph	Left Part	Right part	Edges	Avg $d_{out}$	Avg $d_{in}$	Diameter	Edge Ratio
CALL	IN-IN	101367	-	113541	1.1201	1.1201	4	0.13X
	IN-SCC	331109 (IN)	502152 (SCC)	877417	2.64993	1.74731	1	X
	IN-OUT	70083 (IN)	145850 (OUT)	203335	2.90135	1.39414	1	0.23X
	SCC-SCC	1235221	-	4444481	3.59813	3.59813	15	5.06X
	SCC-OUT	500900 (SCC)	478196 (OUT)	861238	1.71938	1.80101	1	0.98X
	OUT-OUT	173089	-	138177	0.7983	0.7983	4	0.16X
SMS	IN-IN	35649	-	35832	1.00513	1.00513	6	0.32X
	IN-SCC	94707 (IN)	75475 (SCC)	110280	1.16443	1.46115	1	X
	IN-OUT	10304 (IN)	11436 (OUT)	12879	1.2499	1.12618	1	0.12X
	SCC-SCC	440085	-	1466544	3.33241	3.33241	16	13.3X
	SCC-OUT	79844 (SCC)	135698 (OUT)	157932	1.97801	1.16385	1	1.43X
	OUT-OUT	73190	-	73610	1.00574	1.00574	5	0.67X

TABLE XVII

LINEAR CORRELATION COEFFICIENTS BETWEEN CALL AND SMS GRAPHS

SMS Graph Variable	Call Graph Variable	Coefficient
<i>Outdegree</i>	<i>Indegree</i>	0.0179568
<i>Outdegree</i>	<i>Outdegree</i>	0.974498
<i>Indegree</i>	<i>Indegree</i>	0.206114
<i>Indegree</i>	<i>Outdegree</i>	0.154815
<i>Pagerank</i>	<i>Pagerank</i>	0.131639

#### D. Correlation between SMS and Call Graphs

We studied if there was some correlation between the in-degree, out-degree and pagerank of nodes in the call graph, and these values for the same nodes in the SMS graph. Table XVII shows the linear correlation coefficient values observed for various pairs of variables studied. For conducting these experiments we segregated the nodes which were common in both the graphs. The number of such nodes was 1035025, which is almost the same as the number of nodes in the SMS graph (1162457). However this number is less than half of the total nodes in the call graph (2324556); as we have already seen, there exist a large number of nodes which are only present in the call graph and not in the SMS graph.

Table XVII shows that the out-degrees of the two graphs are remarkably correlated. This indicates that people who make a lot

of calls also sends many SMSes. However indegree and pagerank do not seem to be correlated. There may be a lot of nodes in the SMS graph which have incoming edges in the call graph, from nodes which are not present in the SMS graph, making them have a high in-degree in the call graph but not the SMS graph. On the other hand some nodes have a high in-degree in the SMS graph but not in the call graph. *This statistics again emphasizes the point made earlier that Telecom Operators need to analyze the SMS graphs also to identify important segments of customers who may not be discovered by call graph analysis.*

## VIII. 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 analyze network topologies. We used a carefully chosen set of parameter which reveal mostly connectivity directed characteristics and used them on Call and SMS graphs of a mobile operator. 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 the telecom graphs and also enable researchers and business community to gain insight into the graphs. These results can significantly affect business strategies. Therefore, at present we are currently collaborating

with the marketing department of the Telecom Service provider to derive deeper business insights from the results.

The shapes 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 from other countries can serve to verify or refute this model. In fact it will be interesting to determine whether the properties of the Call and SMS graphs determined in our study hold in other geographies and cultures as well.

Our analysis indicates that there are some similarities as well differences between the call and SMS graphs. The SMS graph seems to be more *social* with a higher value of reciprocity. We believe that the Telecom operators need to analyze both these graphs to gain complete understanding of its customer base.

Our temporal analysis highlights interesting insights on how the graphs evolve over time. In the future we plan to study the evolution of the graphs for a longer period of time. A problem worthy of consideration is to find a generative model for these telecom graphs, if one exists. If found, it is likely to offer deep insights into how a mobile operator's customer base evolves with time.

## REFERENCES

- [1] J. Abello, P. M. Pardalos, and M. G. C. Resende. On Maximum Clique Problems in Very Large Graphs. In J. Abello and J. Vitter, editors, *External Memory Algorithms (DIMACS Series)*, pages 119–130. American Mathematical Society, 1999.
- [2] William Aiello, Fan Chung, and Linyuan Lu. A Random Graph Model for Massive Graphs. In *Proceedings of the Thirty-second annual ACM symposium on Theory of Computing*, pages 171–180, May 2000.
- [3] Wai Ho Au Chan and K. C. C. Xin Yao. A Novel Evolutionary Data Mining Algorithm with applications to Churn Prediction. *IEEE Transaction on Evolutionary Computation*, 7(6):532–545, Dec 2003.
- [4] A. L. Barabasi and Reka Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, October 1999.
- [5] Sergey Brin and Lawrence Page. The Anatomy of a large-scale Hypertextual Web Search Engine. *Proceedings of the Eighth International Conference on the World Wide Web/Computer Networks*, 30(1–7):107–117, 1998.
- [6] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph Structure in the Web. In *Proceedings of the Ninth International Conference on the World Wide Web/ Computer Networks*, volume 33, pages 309–320, 2000.
- [7] G. Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007.
- [8] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. MEDUSA - New Model of Internet Topology Using k-shell Decomposition. In *Proceedings of the International Workshop and Conference on Network Science (NetSci)*, 2006.
- [9] W. de Nooy, A. Mrvar, and Batagelj. V. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.
- [10] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the Webgraph. *The European Physical Journal B*, 38:239–243, 2004.
- [11] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the Inner Structure of the Web Graph. In *Eighth International Workshop on the Web and Databases*, 2005.
- [12] S. Dorogovtsev and J. Mendes. *Evolution of Networks: From Biological Nets to the Internet and the WWW*. Oxford University Press, 2000.
- [13] T. Euler. Churn Prediction in Telecommunications Using MiningMart. In *Proceedings of the Workshop on Data Mining and Business (DMBiz)*, 2005.
- [14] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-law Relationships of the Internet Topology. In *Proceedings of ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pages 251–262, 1999.
- [15] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Journal of ACM*, volume 46, 1999.
- [16] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and Evolution of Blogspace. *Communications of the ACM*, 47(12):35–39, 2004.
- [17] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and Evolution of Online Social Networks. In *Proceeding of the Twelfth ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, August 2006.
- [18] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for emerging Cyber-communities. In *Proceedings of the Eighth International Conference on World Wide Web*, pages 1481–1493, 1999.
- [19] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, Shrinking diameters and possible explanations. In *Proceeding of the Eleventh ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 177–187, 2005.
- [20] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic Routing in Social Networks. *PNAS*, 102(33):11623–11628, 2005.
- [21] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45:167, 2003.
- [22] A. Ntoulas, J. Cho, and C. Olston. Whats new on the Web? The evolution of Web from a Search Engine perspective. In *Proceedings of the Thirteenth International Conference on World Wide Web*, pages 1–12, 2004.
- [23] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. ANF: a Fast and Scalable Tool for Data Mining in Massive Graphs. In *Proceeding of the Eighth ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 81–90, 2002.
- [24] G. Siganos, S. Tauro, and M. Faloutsos. Jellyfish: A Conceptual Model for the AS Internet Topology. *Journal of Communications and Networks*, 8(3):339–350, September 2006.



**Amit Anil Nanavati** is a Research Staff Member in IBM India Research Lab, New Delhi, India. He graduated from Louisiana State University, Baton Rouge with M.S. (1994) and Ph.D. in Computer Science (Distributed Computing) in 1996, and with a B.E. in Computer Science from M.S. University of Baroda in 1989. Prior to IBM Research, he worked for Netscape Communications (1996-2000). His recent research focus has been on mobile and pervasive computing and information mining and retrieval, as applied to Telecom service provider settings and is currently working on solutions for emerging economies. He is particularly interested in applications of graph theory in various domains. His other interests include speech in mobile and pervasive environments and theory and modelling of systems. Before completing his Ph.D., he spent a summer at the Jet Propulsion Laboratory, Caltech, NASA. He can be reached at [namit@in.ibm.com](mailto:namit@in.ibm.com).



**Rahul Singh** received the B.Tech. degree from the Indian Institute of Technology at Guwahati in 2005. From 2005 to 2006, he worked at the Oracle Development Center in Bangalore, India as member of technical staff. In 2006, he joined IBM India Research Lab at New Delhi as a technical staff member. He is currently in the first year of his PhD in Computer Science at the University of Massachusetts, Amherst. His area of research is machine learning, data mining and applications of the above to large-scale systems. He can be reached

at [rahul@cs.umass.edu](mailto:rahul@cs.umass.edu).



**Dipanjan Chakraborty** is a Research Staff Member at IBM India Research Lab. He received his Ph.D. in Computer Science from University of Maryland, Baltimore County (UMBC) in 2004. His research is in the areas of mobile and pervasive computing environments, next generation network protocols and management, peer-to-peer systems with special interests in the fields of service discovery, information aggregation and composition, ad-hoc/sensor networks and application-centric routing. He is also working in the area of business process management.

His thesis is in the area of service discovery and composition for pervasive environments. He received a fellowship grant from IBM during the 3 years of his Ph.D candidacy. He can be reached at [cdipanjan@in.ibm.com](mailto:cdipanjan@in.ibm.com).



**Siva Gurumurthy** received B. Tech. degree in Computer Science and Engineering from Indian Institute of Technology Guwahati in 2005. He had pursued one year research training from IBM India Research Lab. Currently, he is a second year Masters Student at University of Massachusetts Amherst in the ECE Department. His research interests include large scale data analysis such as web graph, call graph and social network analysis. He can be reached at [shiva@gmail.com](mailto:shiva@gmail.com).



**Koustuv Dasgupta** is a Researcher at the IBM India Research Lab, New Delhi, India. Koustuv received his Ph.D. in Computer Science from the University of Maryland Baltimore County in May, 2003, and his B.Engg. in Computer Science and Engineering from Jadavpur University, India in July 1997. At IBM, for the past four years, he has been working on and leading projects on telecom infrastructure and middleware, including algorithms and architectures for advanced SIP-based presence in converged networks, context-sensitive middleware

technologies, and next-generation IMS-based service delivery platforms. His research interests further span a wide range of networked computer systems including Internet-scale systems, QoS-aware storage systems, wireless sensor networks, and enterprise grids. Koustuv is a member of ACM and IEEE. He can be reached at [kdasgupta@in.ibm.com](mailto:kdasgupta@in.ibm.com).



**Sougata Mukherjea** is a Research Staff Member and Manager of the Telecom Research Innovation Center in IBM India Research Lab. He received his Bachelors from Jadavpur University, Calcutta, MS from Northeastern University, Boston and PhD from Georgia Institute of Technology, Atlanta (all in Computer Science). Before joining IBM, he held research and software architect positions in companies in Silicon Valley (California) including NEC USA, Inktomi and BEA Systems. His research interests include Middleware technologies and its applications

to Telecom, Data Analysis, Information Retrieval and Visualization. He has several patents and publications in reputed Computer Science conferences and journals in these research areas. He can be reached at [smukherj@in.ibm.com](mailto:smukherj@in.ibm.com).



**Anupam Joshi** is a Professor of Computer Science and Electrical Engineering at UMBC. Earlier, he was an Assistant Professor in the CECS department at the University of Missouri, Columbia. He obtained a B. Tech degree in Electrical Engineering from IIT Delhi in 1989, and a Masters and Ph.D. in Computer Science from Purdue University in 1991 and 1993 respectively. His research interests are in the broad area of networked computing and intelligent systems. His primary focus has been on data management and security for mobile, pervasive, and

sensor systems. He has created agent based middleware to support discovery, composition, and secure access of services/data over both infrastructure based and ad-hoc wireless networks, as well as systems that integrate sensors with the grid. He is also interested in Semantic Web, Social Media, and Data/Web Mining, where he has worked on creating personalized and secure web spaces using a combination of agents, policies, and soft computing.

He has published over 150 technical papers, and has obtained research support from NSF, NASA, DARPA, DoD, IBM, LMCO, AetherSystems, HP, AT&T and Intel. He has presented tutorials in conferences, served as guest editor for special issues for VLDB J., Comm. ACM etc., and served as an Associate Editor of IEEE Transactions of Fuzzy Systems from 99-03. He currently serves on the editorial board of the Intl. J. Semantic Web and Information Systems. At UMBC, Joshi teaches courses in Operating Systems, Mobile Computing, Networking, Security, and Web Mining. He is a member of IEEE, IEEE-CS, and ACM. He can be reached at [joshi@cs.umbc.edu](mailto:joshi@cs.umbc.edu).



**Gautam Das** received B. Tech. degree in Computer Science and Engineering from Indian Institute of Technology Guwahati in 2005. From 2005 to 2006, he worked at IBM India Research Lab, New Delhi. Currently, he is a second year Masters Student at University of Massachusetts Amherst in the ECE Department. His research interests include product optimization based on data analytics, social network analysis, mining and characterization of user generated content. He can be reached at [gdas@ecs.umass.edu](mailto:gdas@ecs.umass.edu).