Disentangling Social Networks inferred from Call Logs

Manuel Cebrian and Alex Pentland
The Media Laboratory, Massachusetts Institute of Technology, Cambridge 02139, MA, USA

Scott Kirkpatrick
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Analysis of an unusually detailed telephone call data set — a month of nearly all mobile and landline phone calls placed during August 2005 the United Kingdom — allows us to identify several different types of social networks that are formed, and relate them to different activities that generate them. We distinguish, among others, work-related and personal or leisure-focused activities and show that the networks they form have very different characteristics. Our principal tool for the analysis, k-core decomposition, shows that distinct distributions of connectivity are present in the two spheres, and that this differentiation affects dramatically the dynamics of information diffusion. Both differ from the simpler and more globally connected structure evident in communications data such as the Internet AS graph.

arXiv:1008.1357v1 [cs.CY] 7 Aug 2010

We know from common sense that the social network of a country is composed of multiple entangled networks, or communities, such as work and leisure, which are different, and have different requirements, but these are difficult to disentangle.[1–8]. We present here a new approach to this question.

Our objective is to obtain functional inferences one might want to make from call graphs regarding their structure, with applications to modeling innovation, political opinion-making, emergency notification, advertising and various other sorts of socially relevant communications among friends, family, work groups and the like.

I. DATA TO BE STUDIED

We use a month of nearly all mobile and landline phone calls placed during August 2005 in the United Kingdom to infer its social communication structure. This data set comprises 65×10^6 different numbers (anonymized, with the area prefix and phone number separated, both then hashed) and 368×10^6 phone calls, timestamped to the second, which represents more than 90% of the mobile phones and greater than 99% of the residential and business landlines in the country. We further consider the links (i, j) to be undirected, storing them in a canonical order $(i < j)$.

We track separately the number of calls $i \rightarrow j$ and $j \leftarrow i$, as is customary in discussions of social networks, since highly directional links express a different sort of relationship than links in which the calls are reciprocated relatively equally. This data set compares with others recently published [9–15]. In the process of aggregation, we also save the total duration of calls between each pair of numbers in each direction. Finally, we construct subnetworks which should represent a stronger degree of minimum relationship between nodes by restricting attention to links in which calls are reciprocated – the smaller of the number of calls from i to j and the number of calls from j to i is at least 1 or 2.

Metro area	#nodes	#work links	#leisure links
PnLa	3,000,119	21,471,591	10,683,865
VCWy	1,866,693	11,727,023	3,398,336
GNgr	866,402	6,540,751	3,051,083

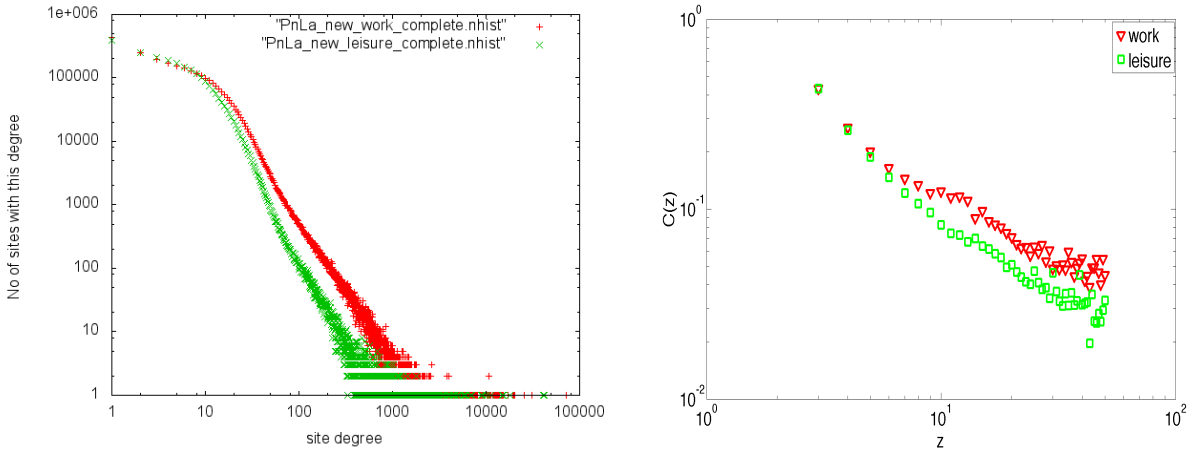
TABLE I: Three large metropolitan areas in the UK.

Given the large amount of work on Internet-derived networks, it is natural to ask if insights from communications networks can be applied to this data, or are social networks different? To answer this we extract three metropolitan areas (phone number prefixes) from the call logs. Within these 3 metropolitan areas we distinguish *work* and *leisure* calls. Work calls are those placed from 8:00 a.m. to 6:00 p.m. during the week. Leisure calls are those placed between 6:00 p.m. and 8:00 a.m. during the week, plus all phone calls on the weekend. August is a popular month for vacations, so this separation is not absolute. Table I summarizes this data.

The results shown throughout the paper are consistent across the 3 metropolitan areas. We restrict our presentation to PnLa in this report.

Figure 1 shows degree and clustering coefficient distributions. They show small differences between work and leisure, but give no insights into their causes. The degree distributions are quite similar, although the leisure network nodes are generally of smaller degree. In fig. 1(b), we also observe small differences in the clustering coefficients. Even though Newman conjectured social networks – as opposed to communication networks – to be assortative [16], our measurements show that social networks inferred from call logs are dissasortative, which Ravasz and Barabasi and have attributed to an underlying hierarchical structure [17].

A larger variation is seen between work and leisure in a simple model of overall diffusion of information (see Fig. 2). We consider how information diffuses out from low degree sites, using the call network in PnLa, restricted to links which are reciprocated at least once. In our model,



(a)Degree. (b)Clustering coefficient.

FIG. 1: PnLa's networks of work and leisure.

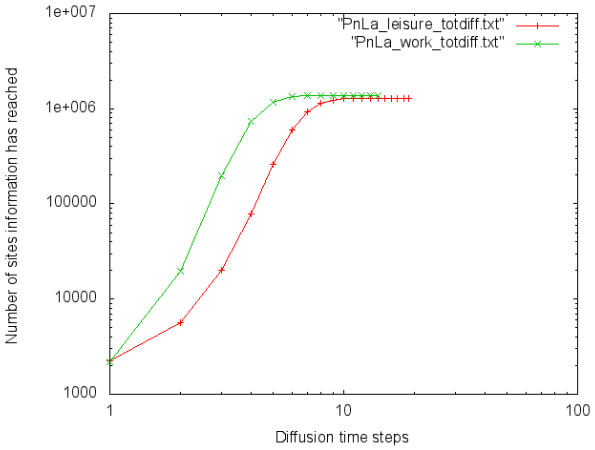


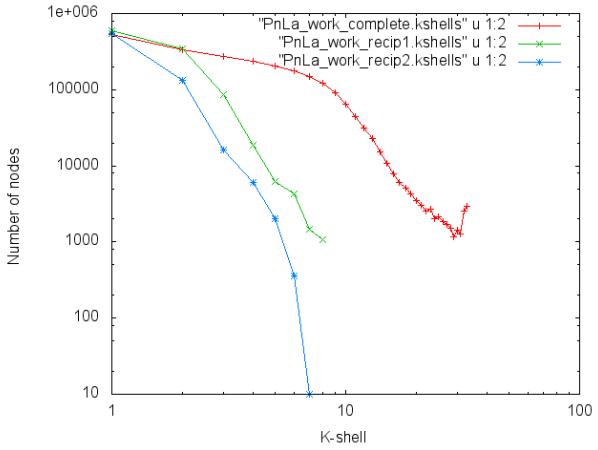
FIG. 2: Simple diffusion rate on work and leisure networks.

information takes one hop to move from one site to all of its neighbors. The time axis is number of hops, and the vertical axis is the number of nodes which the information has reached. The work network spreads information roughly twice as fast as does the leisure network in this simple model. These differences, however, cannot be understood without extracting more local characteristics of the processes involved.

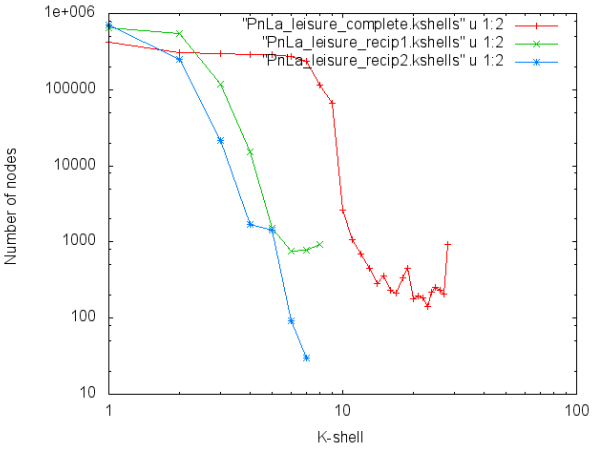
We construct our analysis in two steps, first using k-core decomposition into k-shells to relate the role of each node (here a phone number) to its local environment in the call graph. Then we run a simple diffusion model of information spreading in the call graph to relate function to our k-shell labelling of the nodes in the graph. We describe the decomposition in detail below. It has been used in the analysis of communication networks, and identifies a simple and elegant structure for the graph of the Internet's constituent subnetworks, or autonomous systems (ASes) [18, 19]. As we will illustrate, our call networks display a much more complex structure with evident differences between the leisure and work time period networks, when disentangled in this way. A greater fraction of the nodes form a loose edge of small clusters of individuals, poorly connected over greater scales. Some of the fractal features seen in communications networks are also seen in these social networks, but with very different exponents, which are affected by the difference between work and leisure connectivity. One characteristic that is common to both social and communications networks is the existence of a central nucleus, which is very highly and robustly connected. We show next how this can be uniquely defined.

The k-core is an old concept from graph theory [20]. It is the largest subgraph of selected sites and their direct interconnections such that all sites have degree $>= k$. A k-core is unique and easily found. The usual procedure is to first prune, recursively, all sites with only one neighbor, removing that link as well, until all sites remaining have degree two or more. The sites removed at this stage are called the 1-shell, and what remains is the 2-core. In this way a series of k-shells are identified and removed, leaving at each step a k+1-core. The union of shells 1 through k constitutes a "k-crust." K-crusts, shells, and cores have been studied in random graphs and in communication graphs [18, 19, 21–23].

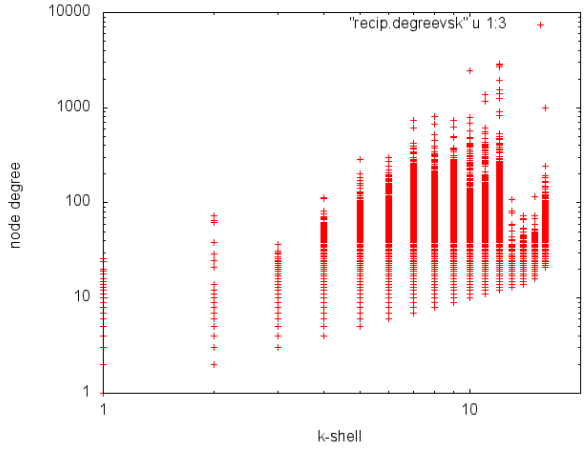
There are some surprising properties that this decomposition can expose. Bollobás showed that in Erdos-Renyi style random graphs, each k-core is with high probability k-connected. This means that every pair of sites is joined by k or more entirely disjoint paths, with no intermediate link or site in common. This is a very strong demonstration of robustness for whatever information the network distributes. Second, Shalit et al.[21] showed that for a simple class of random networks with a long-tailed degree distribution, the sizes of the k-cores follow a power law distribution, decreasing ap-



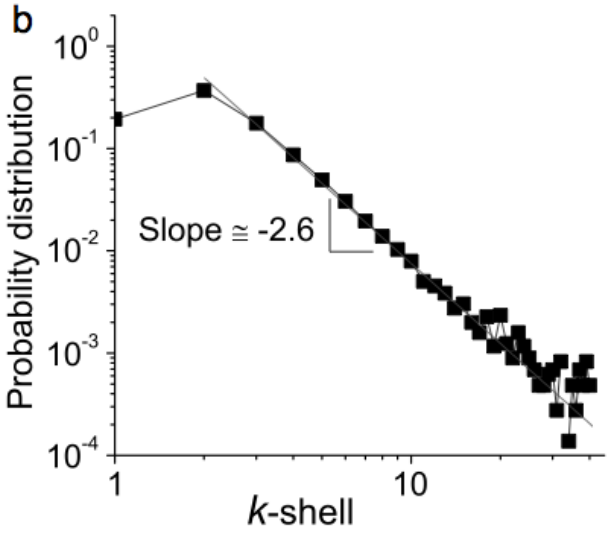
(a)Work.



(b)Leisure.



(c)Average degree of the sites in each shell (work and leisure networks).



(d)The contribution of each shell to the peer-connected component in the Internet AS system (extracted from [19]).

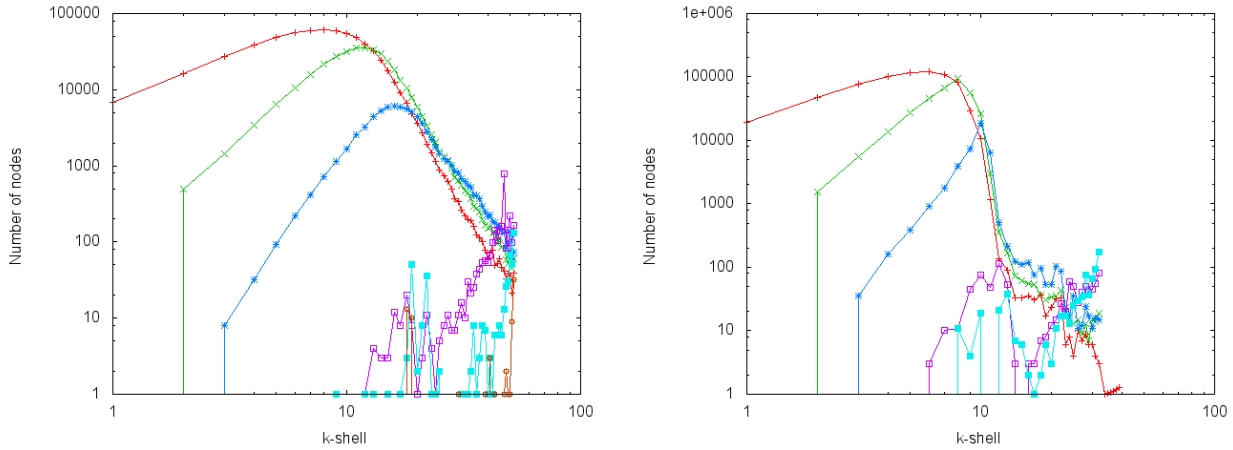
FIG. 3: K-core decomposition of PnLa.

proximately as k^{α} with$-2.4 \leq \alpha \leq -2.8$ for the shells, and $-1.4 \leq \alpha \leq -1.6$ for the cores. The k-shell distribution ends at a value k_{max}. For the random networks, this is simply the point at which k exceeds the number of sites remaining, but for communications networks, the K_{max}-core contains many more sites than k_{max}, and forms a natural "nucleus" of the network. It contains the sites and links with highest betweenness centrality. In the actual Internet AS-graph, the nucleus or k_{max}-core consists of major international and large country carriers plus a few companies, such as Google, which have created their own multi-continent data networks.

In this work, we discover that the telephone call-networks, analyzed using k-core decomposition, have a number of characteristics that are different. In particular, the earlier k-shells of the call-network are more complex, and seem to capture the rich local structure of a society. This frames the question of whether a social network, grown bottom-up from local relationships, is fundamentally different from a communications network, which is strongly shaped by top-down design principles and the objective to provide global communications paths connecting all participants.

II. K-SHELL DECOMPOSITION

To understand better the local structure of these networks, we now construct their k-shell decompositions. The results are shown in 3(a) and 3(b), first for the work links, and then for the leisure network. In each figure we first show the number of nodes in each of the complete subnetwork's k-shells, followed by the same information for the k-shells of the restricted networks in which we



(a)Work. (b)Leisure.

FIG. 4: PnLa's contributions of the doubly reciprocated call graph by k-shell to the k-shells of the complete call graph using work and leisure data. The upper three curves show the contributions of k-shells 1, 2, and 3 in the doubly-reciprocated call-network during work and leisure hours to various k-shells of the full call graph. The lower three curves show the contributions of shells 7, 8, and 9.

include only singly- and doubly-reciprocated links, those for which the number of calls in the less-common direction is at least 1 or 2. Note that separating the nodes into their k-shells corresponds only roughly to an ordering by node degree, as shown in the third section of this figure. Finally, for comparison, we include the result from Carmi et al. for the Internet AS-graph as Fig. 3(d). The k-shell size distributions for the call graphs are very different in shape from that of an Internet graph, with the flat initial sections of the complete graphs (not restricted to reciprocated links) extended over 10 shells, rather than one or two shells. The slopes of the descending parts of both curves are steeper than is seen in Fig. 3(d). But both graphs have clearly identifiable nuclei in the complete and singly reciprocated networks.

Fig. 3 also shows differences between work and leisure k-shells. There are similarities in the initial shells, but much more rapid decrease in the "fractal" part of the curve. As in Fig. 3d, there is a big jump to the nucleus, but the size of the nucleus is over 1000 sites in both call graph networks, and less than 100 in the AS graph. The plots show that the deepest shell of the reciprocal network contributes mostly to the deepest shell of the overall network, but only a little to the second peak, which may be a different cluster.

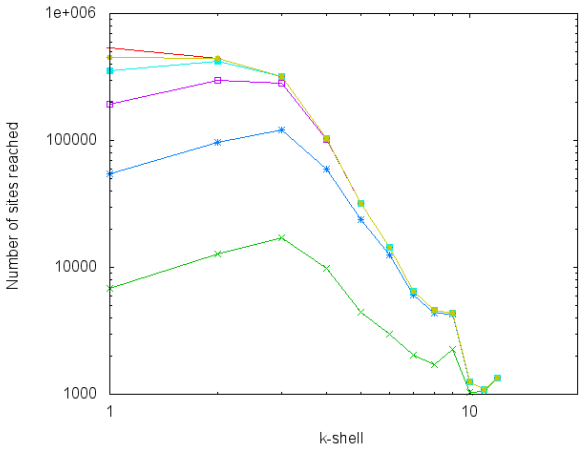
Note that in Fig. 3(c), while the average degree of the sites in each shell increases monotonically, there is an exceptional behavior near the end of the distribution, with the highest degree nodes appearing not in the nucleus shell, but a few shells earlier, followed by a small set of sites of large but not so extreme degree. This remains unexplained at present.

We plot in Fig. 4 the contributions from each k-shell of the very restricted subnetworks in which all calls are reciprocated at least twice to the k-shells of the full call graph. Again, we separate work and leisure time periods. The earlier k-shells of the restricted subnetwork contribute to the outer half of the k-shells in the full network, with the largest contributions seen in the shells of the full network where the power law-like decrease finally begins. The contribution of the reciprocated calls to the outer k-shells of the full call graph is quite negligible. However, the inner k-shells of the full call graph are made primarily of the innermost shells of the reciprocated subnetwork.

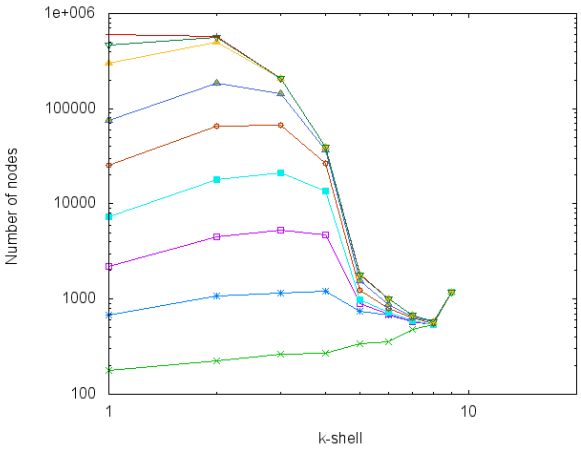
III. INFORMATION DIFFUSION

Now we move on to understanding the different ways in which information diffuses in the work and the leisure networks. The traditional approach might be to note how rapidly information spread from the sites of lower degree reaches the whole network. We show this in fig 2. There is a clear difference between the slower initial spread within the leisure network and the more rapid initial diffusion in the work network but both conclude the process at a similar rate. We need more information about the networks' local structure to understand the differences. Analyzing diffusion patterns from these different cores and shells reinforces the k-shell analysis, and actually makes these structures more evident. A preliminary examination indicates that the differences between information diffusion in the work and the leisure networks lies in the different properties of the outer shells, as well as overall greater connectivity of the work network compared with the sparser leisure network.

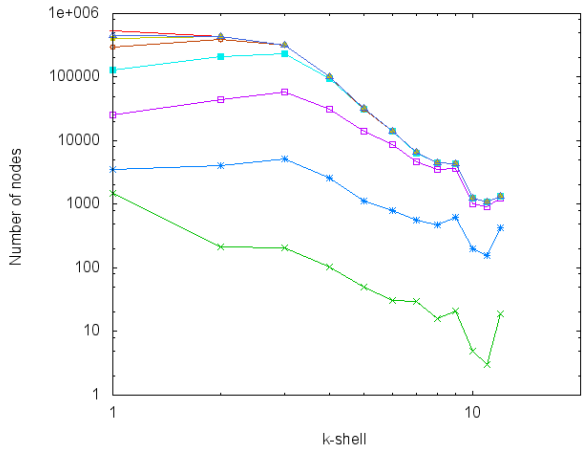
Each curve in Figs. 5 shows the number of sites in each k-shell of the singly reciprocated subnetworks for the work and leisure periods in PnLa's data that have



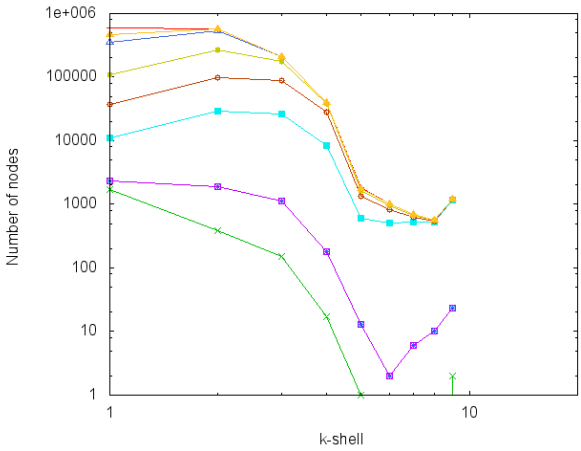
(a)Diffusion from the Kernel in the work network.



(b)Diffusion from the Kernel in the leisure network.



(c)Diffusion from the $K = 1$ shell in the work network.



(d)Diffusion from the $K = 1$ shell in the leisure network.

FIG. 5: Diffusion from the kernel and outer shell for both work and leisure networks. The lines indicate, from top to bottom, the spread at times 1, 2, 3, 4, 6 and 10 time steps. The upper line is the number of sites in each shell. For this study, the network was restricted to links with at least one reciprocal call.

received information after a given number of iterations. In Figs. 5(a) and 5(b) the network was initialized with all sites in the k_{max} shell having information, which then diffuses by direct links at each time step. Information diffuses relatively quickly in the inner shells of both networks, reaching all sites in the inner shells in two iterations[24]. Diffusion into the outermost two shells is slower in the work network of Fig 5(a) and 5(c), and is slower into the outermost three shells in the leisure network, as seen in Fig 5(b) and 5(c). In general, diffusion of information among the sites in the outer shells is about a factor of two times slower in the leisure network than in the work network. Notice also that a few sites in the $k = 1$ shell are never reached by the diffusing information, because they are simply not connected with the rest.

We would also like to know how diffusion processes proceed when starting from a general site in the network. Since most of the sites lie in the outermost two or three shells, we initialized our model with $1,000$ sites labelled in the k=1 shell of each network (leisure in Fig 5(c), work in Fig 5(d)). For the first two iterations, information slowly makes its way to the nucleus sites. Once most of the nucleus sites have received our diffusing information, they take over the principal role in further spread of information, quickly saturating the center of the network, while diffusion proceeds pretty much as in the first experiments at the edges of the network. Once again the tighter and wider-reaching connectivity of the work network makes the process conclude approximately twice as fast as in the leisure network. In constructing the total diffusion curves of Fig. 2, we also followed this procedure, using $1,000$ sites in the $k = 1$ shell of the total network for PnLa (work and leisure combined) as the sources. This indicates that that work and leisure activities have very different characteristics. This may be a function of their different requirements, or show that a much stronger hierarchy organizes communications in

the work network.

There is more to be learned about the local structure of the outer shells. In the analysis of communication networks it is important to understand the breakdown of the network into connected components, and to refine this analysis to regions such as cores and crusts. Communications networks are connected overall, but the crusts exhibit a percolation transition, with a large fraction of sites remaining isolated from the "infinite cluster" until the nucleus is added, to link these with the rest of their network. We have not yet performed the comparable analysis for the call graphs, but expect to have this done by the time of the 2010 WIN Workshop. In particular, we suspect that a structure of many disconnected local islands persists throughout the initial flat regions of the k-shell size curves in Fig. 3, since the first and second shells in communications networks, the ones which depart from a nearly power law decrease in size with increasing k, are so constructed.

Does this pattern emerge even more strongly in the data for the whole country (or larger regions)? We are in the process of aggregating and reducing the data for the whole country and will report on that much larger data set at a later time.

How can we best use this information? Does it emerge quickly or take time to be evident as we accumulate observations of a social network? In terms of call graph data such as we are studying, do the patterns emerge across a country in days, weeks or months? Uses of such data to understand the spread of opinion within a country will require understanding these time constants.

[1] M. Girvan and M. Newman, Proceedings of the National Academy of Sciences **99**, 7821 (2002).

[2] I. Derényi, G. Palla, and T. Vicsek, Physical review letters **94**, 160202 (2005).

[3] J. Leskovec, K. Lang, and M. Mahoney, in *Proceedings of the 19th international conference on World wide web* (ACM, 2010), pp. 631–640.

[4] J. Sun, C. Faloutsos, S. Papadimitriou, and P. Yu, in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2007), p. 696.

[5] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature **435**, 814 (2005).

[6] J. Leskovec and E. Horvitz, in *Proceeding of the 17th international conference on World Wide Web* (ACM, 2008), pp. 915–924.

[7] Y. Ahn, J. Bagrow, and S. Lehmann, Nature **446**, 761 (2009).

[8] P. Mucha, T. Richardson, K. Macon, M. Porter, and J. Onnela, Science **328**, 876 (2010).

[9] N. Eagle, M. Macy, and R. Claxton, Science **328**, 1029 (2010).

[10] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. Barabási, Proceedings of the National Academy of Sciences **104**, 7332 (2007).

[11] P. Wang, M. Gonzalez, C. Hidalgo, and A. Barabasi, Science **324**, 1071 (2009).

[12] M. Cebrian, M. Lahiri, N. Oliver, and A. Pentland, IEEE Journal of Selected Topics in Signal Processing **4**, 677 (2010).

[13] G. Krings, F. Calabrese, C. Ratti, and V. Blondel, Journal of Statistical Mechanics: Theory and Experiment **2009**, L07003 (2009).

[14] N. Du, C. Faloutsos, B. Wang, and L. Akoglu, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2009), pp. 269–278.

[15] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2008), pp. 596–604.

[16] M. Newman, Physical Review Letters **89**, 208701 (2002).

[17] E. Ravasz and A. Barabási, Physical Review E **67**, 26112 (2003).

[18] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, Arxiv preprint cond-mat/0601240 (2007).

[19] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, Proceedings of the National Academy of Sciences **104**, 11150 (2007).

[20] B. Bollobás, *Random graphs* (Academic Press, 1985).

[21] A. Shalit, S. Kirkpatrick, and S. Solomon, in *Aspects of Complexity and Its Applications, Rome* (2002).

[22] S. Dorogovtsev, A. Goltsev, and J. Mendes, Physical review letters **96**, 40601 (2006).

[23] J. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, Networks and Heterogeneous Media **3** (2008).

[24] In Carmi et al. [19], the k-core analysis of the AS graph showed that cores well outside the nucleus still had diameter of two, so information can saturate those inner shells in two hops, as we see in both Fig 5(a) and 5(b)).