New Journal of Physics
The open–access journal for physics

Analysis of a large-scale weighted network of one-to-one human communication

Jukka-Pekka Onnela[1,2], Jari Saramäki[1], Jörkki Hyvönen[1], Gábor Szabó[3,4], M Argollo de Menezes[3], Kimmo Kaski[1], Albert-László Barabási[3,4] and János Kertész[1,5]

[1] Laboratory of Computational Engineering, Helsinki University of Technology, Finland
[2] Clarendon Laboratory, Physics Department, Oxford University, Oxford, UK
[3] Department of Physdics and Center for Complex Networks Research, University of Notre Dame, IN, USA
[4] Center for Cancer Systems Biology, Dana Farber Cancer Institute, Harvard University, Boston, MA, USA
[5] Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary
E-mail: kertesz@phy.bme.hu

New Journal of Physics **9** (2007) 179
Received 13 February 2007
Published 28 June 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/6/179

Abstract. We construct a connected network of 3.9 million nodes from mobile phone call records, which can be regarded as a proxy for the underlying human communication network at the societal level. We assign two weights on each edge to reflect the strength of social interaction, which are the aggregate call duration and the cumulative number of calls placed between the individuals over a period of 18 weeks. We present a detailed analysis of this weighted network by examining its degree, strength, and weight distributions, as well as its topological assortativity and weighted assortativity, clustering and weighted clustering, together with correlations between these quantities. We give an account of motif intensity and coherence distributions and compare them to a randomized reference system. We also use the concept of link overlap to measure the number of common neighbours any two adjacent nodes have, which serves as a useful local measure for identifying the interconnectedness of communities. We report a positive correlation between the overlap and weight of a link, thus providing

PII: S1367-2630(07)43983-0

strong quantitative evidence for the weak ties hypothesis, a central concept in social network analysis. The percolation properties of the network are found to depend on the type and order of removed links, and they can help understand how the local structure of the network manifests itself at the global level. We hope that our results will contribute to modelling weighted large-scale social networks, and believe that the systematic approach followed here can be adopted to study other weighted networks.

Contents

1. Introduction and data

Social networks have been a subject of intensive study since the 1930s. In this framework social life consists of the flow and exchange of norms, values, ideas and other social and cultural resources [1], and social action of individuals is affected by the structure of the underlying network [2]. The structure of social networks is important then not only from the perspective of the individual, but also from that of the society as a whole. However, uncovering the structure of social networks has been constrained by the practical difficulty of mapping out interactions among a large number of individuals. Social scientists have ordinarily based their studies on questionnaire data, typically reaching the order of $N \approx 10^2$ individuals [3]. Although the spectrum of social interactions that may be probed in this approach is wide, the strength of an interaction is often based on recollection and, consequently, is highly subjective. However, in the late the 1990s a change of paradigm took place [4, 5]. Physicists became interested in large-scale social networks, utilizing electronic databases from emails [6]–[8] to phone records [9], offering unprecedented opportunities to uncover and explore large-scale social networks [10]. In this scheme the order of $N \approx 10^6$ individuals may be handled and, although the range of social interactions is narrower, in some cases their strengths may be objectively quantifiable. While both approaches have their merits, studying large-scale networks has potential to shed light on how individual microscopic interactions translate into macroscopic social systems. In addition to this being one of the key questions as posed by social scientists in the field, it is also the one to which statistical physics in general, and the science of complex networks in particular, can make a contribution.

In this paper, we present a detailed analysis of a network constructed from a data set consisting of the mobile phone call records of over seven million individuals over a period of 18 weeks (126 days), covering approximately 20% of the population of the country. For the

purpose of retaining customer anonymity, each subscription was identified by a surrogate key, guaranteeing that the privacy of customers was respected. We kept only voice calls, filtering out all other services, such as voice mail, data calls, text messages, chat and operator calls. We filtered out calls involving other operators, incoming or outgoing, keeping only those transactions in which the calling and receiving subscription is governed by the same operator. This filtering was needed to eliminate the bias between this operator and other operators as we have a full access to the call records of this operator, but only partial access to the calls made to subscriptions governed by other operators. We constructed two different networks from the data. In the first scheme we connected two users with an undirected link if there had been at least one phone call between them, i.e. i called j *or* j called i, resulting in a *non-mutual network* consisting of $N = 7.2 \times 10^6$ nodes and $L = 22.6 \times 10^6$ links. However, many of these calls are one-way, most of which correspond to single events, suggesting that they typically reach individuals that the caller might not know personally. To eliminate them, in the second scheme we connected two users with an undirected link if there had been at least one reciprocated pair of phone calls between them, i.e. i called j *and* j called i, resulting in a *mutual network* with $N = 4.6 \times 10^6$ nodes and $L = 7.0 \times 10^6$ links.

The resulting mobile call graph (MCG) naturally captures only a sub-set of the underlying social network, which consists of all forms of social interactions, including face-to-face interactions, email and landline communication etc. However, research on media multiplexity suggests that the use of one medium for communication between two people implies communication via other means as well [11]. Furthermore, in the absence of directory listings, the mobile phone data is skewed towards trusted interactions, i.e. people tend to share their mobile numbers only with individuals they trust. Therefore, the MCG can be used as a proxy for the underlying social network.

We can quantify the weight of the link (i, j) by the aggregated time i and j spent talking to each other as well as by the total number of calls made between i and j over the studied period. These weights are denoted by w_{ij}^D (total duration of calls) and w_{ij}^N (total number of calls), respectively, where the former is measured in seconds (s) and the latter is a dimensionless quantity.

This paper is devoted to the study of these weighted, large-scale, one-to-one social interaction networks, with emphasis on the mutual over the non-mutual network. We adopt a 'cookbook approach' by carrying out a systematic analysis of basic and more advanced network characteristics, and hope that others working on weighted networks will benefit from our 'recipes'. We study some of the basic network characteristics in section 2 and focus on weighted network characteristics in section 3. We explore the coupling between link weight and the surrounding local network topology in section 4. We have dedicated section 5 to the study of percolation properties of the network and, finally, discuss our findings in section 6.

2. Basic network characteristics

We start inspecting the network by showing a small sample of it in figure 1. The sample has been extracted from the mutual network by picking a node (source node) at random and including all nodes in the sample that are within a (topological) distance of $\ell = 5$ from the source node. This method of sampling is sometimes called snowball sampling [12]. The colour of links corresponds to the strength of each tie in terms of w_{ij}^D. It appears from this figure that the network consists

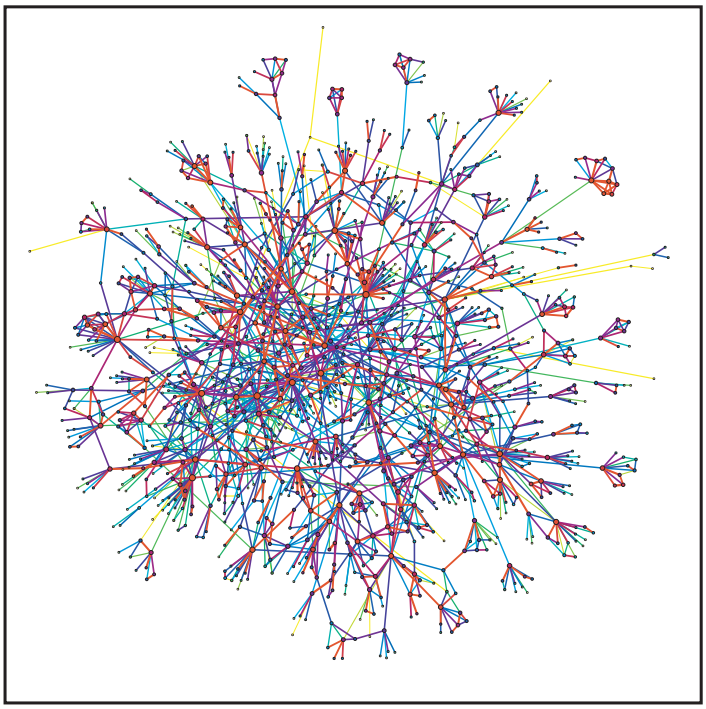

Figure 1. A small sample of the network with link weights w_{ij}^{D} colour coded from yellow (weak link) to red (strong link).

of small local clusters, and the majority of the strong ties (coloured in red) seem to be localized within these clusters. In some cases nodes connected by a strong link have many common neighbours, but there are also strongly connected nodes with few or no common neighbours.

These two apparently contradictory trends arise as a result of being forced to examine a sample of the network as opposed to the entire network. To understand the limits of visual inspection, it is important to realize that since the network is a high dimensional object, a majority of the nodes will be on the outskirts of the sample. A consequence of this is that for most of these nodes we only have partial visibility into their neighbourhood. Consequently, one can see the full neighbourhood for only a small minority of nodes in the sample.

Let us elaborate on network sampling. We show in figure 2 the number of nodes in the sample $N_{\rm s}(\ell)$, obtained using snowball sampling, as a function of extraction distance ℓ for several choices of the source node (solid lines) and their average (dashed line). Here $N_{\rm s}(\ell)$ is defined as the number of nodes within a distance ℓ from the given source node. For a fixed value of x, we call nodes for which $\ell < x$ *bulk nodes* and those with $\ell = x$ *surface nodes* of the sample. The number of surface nodes clearly outweighs the number of bulk nodes. This is to be expected since the network behaves like a high dimensional hypersphere, the volume of which is negligible to its surface area. To a good approximation we can write $N_{\rm s} = A{\rm e}^{B\ell}$, where A and B are fitting parameters. In general, the number of surface nodes to the number of bulk nodes in the sample is $[N_{\rm s}(\ell) - N_{\rm s}(\ell - 1)]/N_{\rm s}(\ell) = 1 - {\rm e}^{-B}$. Thus, assuming that B is of the order of one or smaller, a large majority of nodes in the sample are surface nodes.

This is clear from another network sample in figure 3, in which bulk nodes and surface nodes are drawn with different markers. It is only for bulk nodes that we have a full visibility

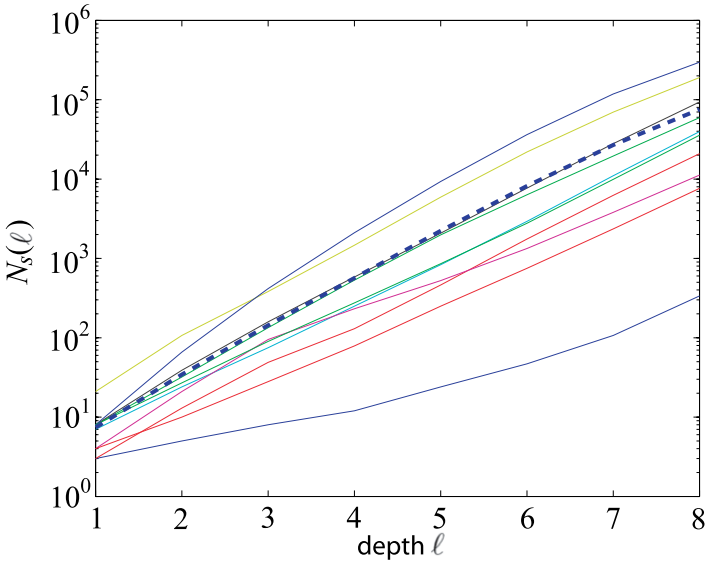

Figure 2. Number of nodes in the sample $N_s(\ell)$, obtained by snowball sampling, as a function of extraction distance ℓ for several choices of the source node (solid lines) and their average (dashed line).

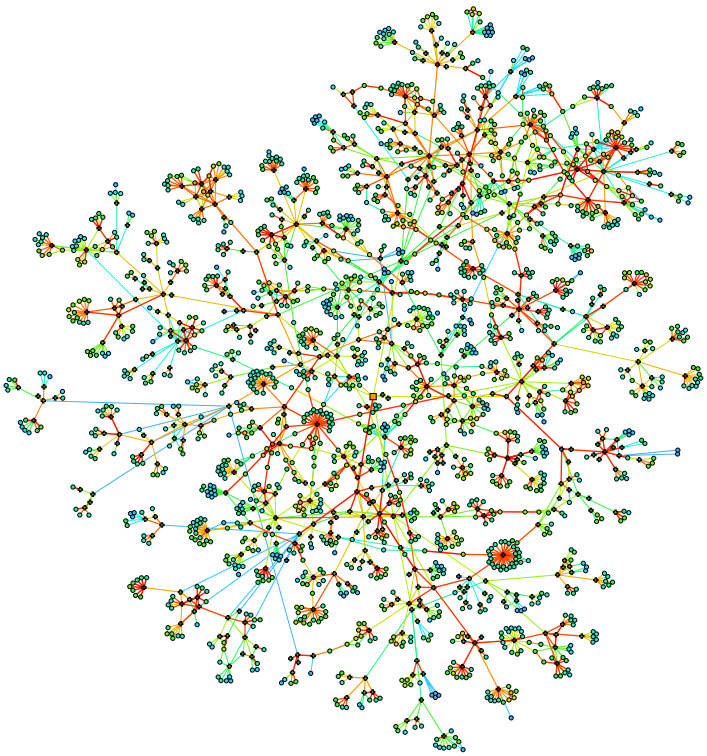

Figure 3. A sample of the network, showing the source (orange square at the centre) node from which sampling was started, the bulk nodes (+), and surface nodes (○). For surface nodes, which clearly are in the majority, only some of their nearest neighbours are visible in the sample, while the rest are outside the sample.

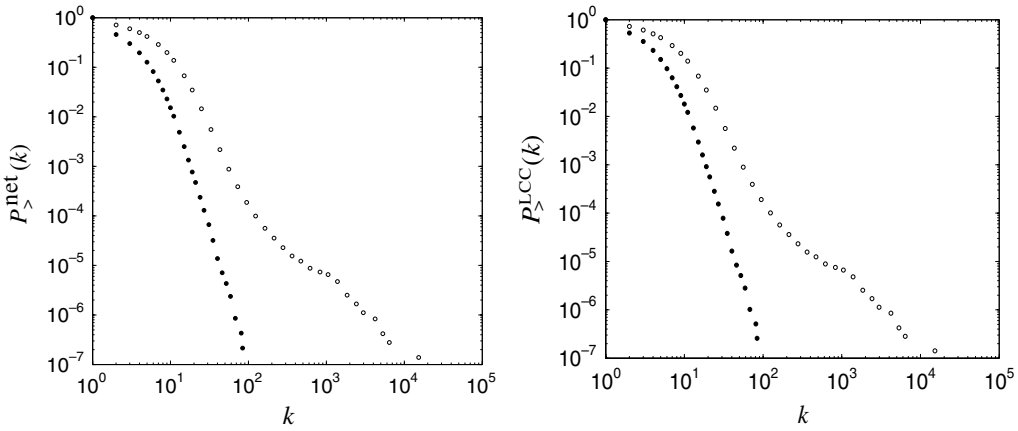

Figure 4. Cumulative degree distribution $P_>^{net}(k)$ for the mutual (•) and non-mutual (○) networks (left) and for their respective largest connected components (LCC) $P_>^{LCC}(k)$ (right). The mutual network is a subgraph of the non-mutual one, and 84.1% of the nodes in the mutual network belong to a single connected component (LCC), for which the average degree $\langle k \rangle \approx 3.0$.

to their neighbourhood and, consequently, may make unbiased judgments about the structure of their neighbourhood. Since these nodes are clearly in the minority, visual inspection of network samples has limited utility.

A basic network characteristic, the degree distribution, is shown in figure 4. To avoid the need for binning, we study the cumulative degree distribution $P_>(k)$, defined as $P_>(k) = \int_k^\infty p(x)\,\mathrm{d}x$, where $p(x)$ the degree probability density function. We denote the distribution for whole mutual and non-mutual networks $P_>^{net}(k)$, and that of their respective *largest connected components* (LCC) by $P_>^{LCC}(k)$. Note that the mutual network is a subgraph of the non-mutual one, and the LCC is a subgraph of the whole network. In the case of the mutual network 84% of the nodes belong to the LCC. In this case little is left outside the LCC, partly explaining why distributions are almost identical for the whole network and the LCC.

In general, the degree distributions are skewed with a fat tail, indicating that while most users communicate with only a few individuals, a small minority talks with dozens. The noticeable difference between the degree distributions for the mutual and non-mutual network is the fatter tail of the non-mutual network. In particular, while the most connected node in the LCC of the mutual network has $k_{max} = 144$, in the LCC of the non-mutual network we find that $k_{max} = 34\,625$. Clearly, the latter cannot correspond to a single individual. However, it appears plausible that the mutual network is dominated by trusted interactions, i.e. people tend to share their mobile numbers only with individuals they trust. We also point out that $k_{max} = 144$ in the mutual network is very close to the approximate number of 150 put forward by Dunbar as a limit on connectivity resulting from the size of neocortex in the cerebral cortex in primates [13]. From now on, unless otherwise mentioned, we shall focus exclusively on the LCC of the mutual network.

The tail of the degree distribution $P(k)$ for the LCC of the mutual network is approximated well by a power law of the form $P(k) = a(k + k_0)^{-\gamma}$ with $k_0 = 10.9$ and $\gamma = 8.4$. Note that the value of the exponent is significantly higher than the value observed for landlines ($\gamma = 2.1$ for

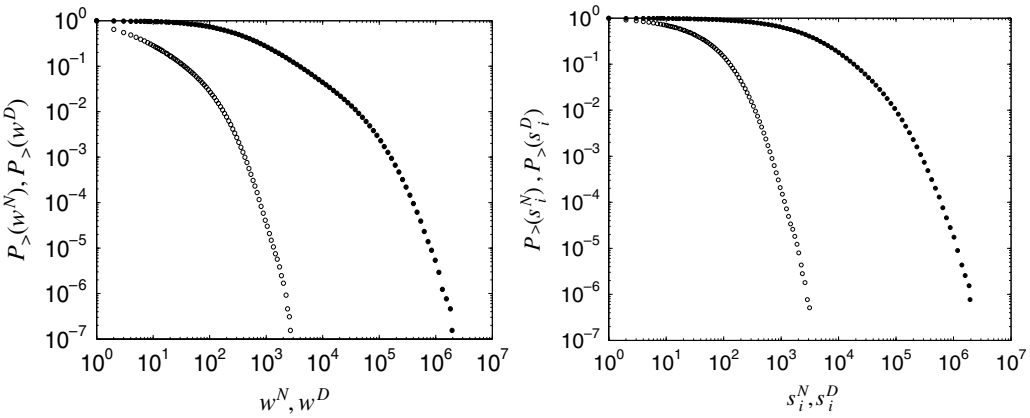

Figure 5. Cumulative link weight distributions (left) and cumulative node strength distributions (right) in the LCC of the mutual network. Link weights and node strengths are measured in terms of the absolute number of calls made during the studied period (○), corresponding to $P_>(w^N)$ and $P_>(s^N)$, as well as the aggregated call duration during the period (●), given by $P_>(w^D)$ and $P_>(s^D)$.

the in-degree distribution [14]). For such a rapidly decaying degree distribution the hubs are few, and therefore many properties of traditional scale-free networks, from anomalous diffusion [15] to error tolerance [16], are absent.

As mentioned in the introduction, the strengths of social interactions are measured both in terms of the aggregate number of calls made, denoted with w_{ij}^N, and the aggregate duration of calls, denoted with w_{ij}^D. The same applies to node strengths, defined as $s_i^N = \sum_{j\in\mathcal{N}(v_i)} w_{ij}^N$ and $s_i^D = \sum_{j\in\mathcal{N}(v_i)} w_{ij}^D$, which correspond to number-of-calls and aggregate-call-duration based strengths, respectively, and where $\mathcal{N}(v_i)$ denotes the neighbourhood of node i. The associated cumulative link weight distributions are $P_>(w^N)$ and $P_>(w^D)$ and the cumulative node strength distributions are $P_>(s^N)$ and $P_>(s^D)$ as shown in figure 5. Both link weight distributions are broad so that while the majority of ties correspond to a couple of calls and a few minutes of air time, a small fraction of users place numerous calls and spend hours chatting with each other. On average an individual made $\langle s^N\rangle \approx 51.1$ calls and spent $\langle s^D\rangle \approx 8074$ s (135 min) on the phone. Two connected individuals spoke on average $\langle w^N\rangle \approx 15.4$ times on the phone spending altogether $\langle w^D\rangle \approx 2429$ s (40 min) talking to one other. These values are summarized in table 1, which also lists some higher moments for the distributions.

The two weights w_{ij}^D and w_{ij}^N are strongly correlated as expected, and this is evident in figure 6. In the mutual network Pearson's linear correlation coefficient r between w_{ij}^N and w_{ij}^D is $r(w_{ij}^N, w_{ij}^D) = 0.70$, implying that variance in w_{ij}^N explains some 50% of variance in w_{ij}^D. The relationship between the two link weights can be further characterized by Spearman's rank correlation coefficient ρ, which assesses how well an arbitrary monotonic function describes the relationship between any two variables without assuming a linear relationship between them. In this case we find that $\rho(w_{ij}^N, w_{ij}^D) = 0.96$. Based on the correlation between link weights, one would expect a correlation between node strengths, and we find here that $r(s_i^N, s_i^D) = 0.68$ and

Table 1. Summary of descriptive network statistics. The following terms are used: whole network (net), LCC, non-mutual network (NM) and mutual network (M). The superscripts N and and D refer to number-of-calls and aggregate-call-duration based link weights and node strengths, respectively.

x	Mean	Std	Skewness	Kurtosis	Max
Degree k_i, net, NM	6.28	16.6	1.39×10^3	2.71×10^6	3.46×10^4
Degree k_i, net, M	3.01	2.41	2.40	17.5	144
Degree k_i, LCC, NM	6.37	16.8	1.38×10^3	2.68×10^6	3.46×10^4
Degree k_i, LCC, M	3.32	2.49	2.28	17.0	144
Weight w_{ij}^N	15.4	37.3	8.54	165	3.61×10^3
Weight w_{ij}^D	2.43×10^3	1.23×10^4	25.1	1.52×10^3	2.39×10^6
Strength s_i^N	51.1	74.8	4.30	44.2	3.64×10^3
Strength s_i^D	8.07×10^3	2.32×10^4	13.5	452	2.48×10^6

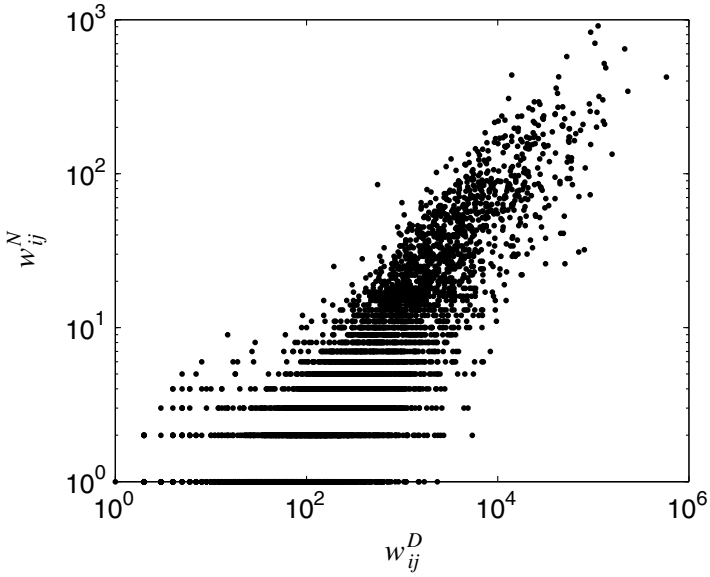

Figure 6. Scatter plot of call duration weights w_{ij}^D and number of calls weights w_{ij}^N. The two weights are clearly correlated in this random sample of 5000 links, as well as in the LCC of the mutual network, giving rise to Pearson's linear correlation coefficient of 0.70 in the latter.

$\rho(s_i^N, s_i^D) = 0.97$. In both cases Spearman's correlation is higher than Pearson's, suggesting that while the associations between w_{ij}^N and w_{ij}^D as well as between s_i^N and s_i^D have linear components to them, the correlations appear to be non-linear in nature[6].

[6] Note that in computing Spearman's rank correlation coefficient we discarded degenerate data points. The degeneracy arises because the data are discrete and its effects are compounded by the fact that the underlying probability distributions are highly skewed. With very large datasets and highly skewed probability distributions this can lead to $\rho \to 0$, giving an erroneous characterization of the data.

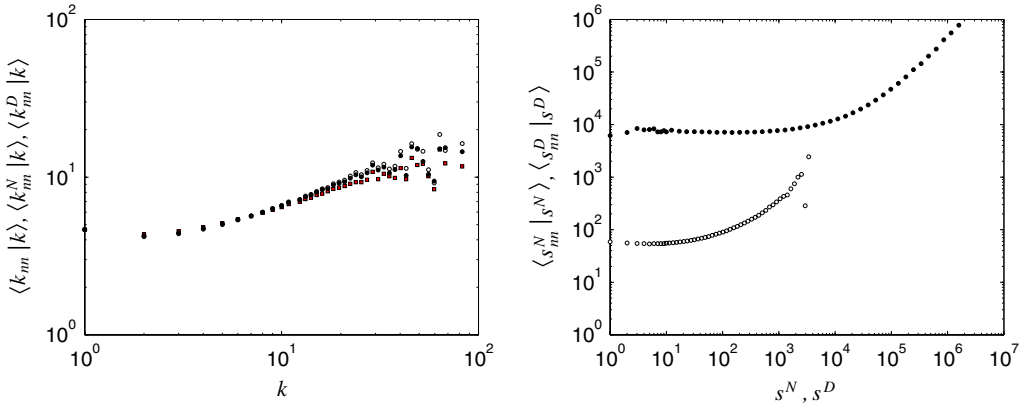

Figure 7. Average neighbour degree $\langle k_{\rm nn}|k\rangle$, $\langle k_{\rm nn}^N|k\rangle$, and $\langle k_{\rm nn}^D|k\rangle$ (left) and average neighbour strength $\langle s_{\rm nn}^D|s^D\rangle$ and $\langle s_{\rm nn}^N|s^N\rangle$ (right) in the LCC of the mutual network. The three markers in the plot on the left correspond to unweighted $\langle k_{\rm nn}\rangle$ (black squares), number-of-calls weighted $\langle k_{\rm nn}^N\rangle$ (○), and call-duration weighted $\langle k_{\rm nn}^D\rangle$ (●) averages. The markers on the right correspond to number of calls (○) and total call duration (●).

The tail of the weight distribution $P(w^D)$ for the LCC of the mutual network is approximated well by an exponentially truncated power-law of the form $P(w) = a(w + w_0)^{-\gamma} \exp(-w/w_{\rm c})$ with $w_0 = 280$, $\beta = 1.9$, and the cutoff parameter $w_{\rm c} = 3.4 \times 10^5$. The broad tailed nature of these distributions is rather unexpected, given that fat tailed tie strength distributions have been observed mainly in networks characterized by global transport processes, such as the number of passengers carried by the airline transportation network [17], the reaction fluxes in metabolic networks [18], and packet transfer on the Internet [19]. In all these cases the individual fluxes are determined by the global network topology, in which an important property is 'conservation of mass', i.e. local conservation of passengers, molecules, and data packets. Such constraints are not present here and, in addition, social networks are expected to be fairly local in nature, meaning that the nature of the link weight and strength distributions are nontrivial. This raises the interesting question of the extent to which network structure and link weights are correlated in this network and, in general, whether their extent of correlation can be used to categorize networks in different classes. We will address the first part of this question in section 4.

Social networks are expected to be assortative: people with many friends are connected to others who also have many friends. This gives rise to degree–degree correlations in the network, meaning that the degrees of two adjacent nodes are not independent. These correlations are completely described by the joint probability distribution $P(k, k')$, giving the probability that a node of degree k is connected to a node of degree k'. It is more practical, however, to define the average nearest neighbours degree of a node v_i as $k_{{\rm nn},i} = (1/k_i) \sum_{j \in \mathcal{N}(v_i)} k_j$, where $\mathcal{N}(v_i)$ denotes the neighbourhood of v_i. By averaging this over all nodes in the network of a given degree k, one can calculate the average degree of nearest neighbours with degree k denoted by $\langle k_{\rm nn}|k\rangle$, which corresponds to $\sum_{k'} k' P(k'|k)$ [20]. The network is said to exhibit *assortative mixing* if $\langle k_{\rm nn}|k\rangle$ increases and *disassortative mixing* if it decreases as a function of k [21].

We show the average nearest neighbour degree in figure 7. We follow Barrat *et al* [22] and use the weighted average nearest neighbour degree to characterize degree–degree correlations, which

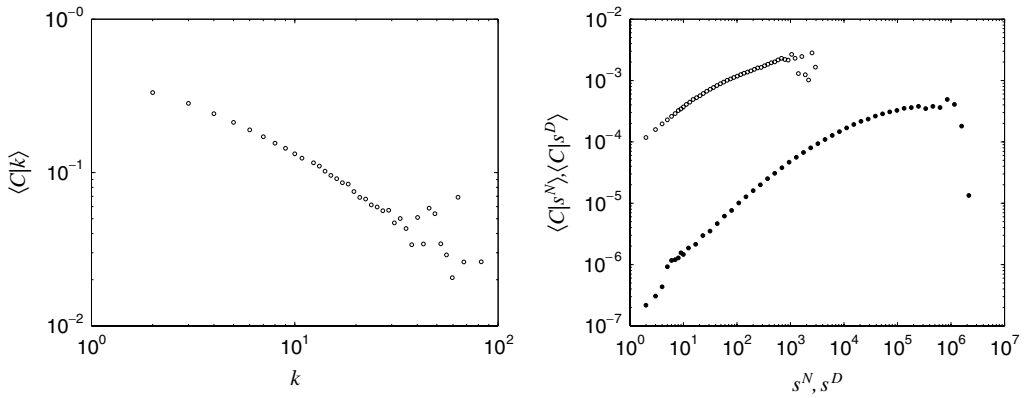

Figure 8. Average (topological) clustering coefficient $\langle C|k\rangle$ (left) and average weighted clustering coefficients $\langle \tilde{C}|s^N\rangle$ and $\langle \tilde{C}|s^D\rangle$ (right) in the LCC of the mutual network. The topological clustering coefficient does not depend on weights, and is presented as a function of degree k (○). In contrast, the weighted clustering coefficient is presented as a function of node strengths in terms of number of calls s^N (○) and aggregated call duration s^D (●).

are written as $k^N_{\mathrm{nn},i} = (1/s^N_i)\sum_{j\in\mathcal{N}(v_i)} w^N_{ij}k_j$ and $k^D_{\mathrm{nn},i} = (1/s^D_i)\sum_{j\in\mathcal{N}(v_i)} w^D_{ij}k_j$, corresponding to the two weighting schemes. Averaging these over the network gives $\langle k_{\mathrm{nn}}|k\rangle$, $\langle k^N_{\mathrm{nn}}|k\rangle$ and $\langle k^D_{\mathrm{nn}}|k\rangle$, which measure the effective affinity to connect with neighbours of a given degree while taking the magnitude of the interactions into account [22]. The three measures behave very similarly in figure 7, and the network is clearly assortative degree-wise such that $\langle k_{\mathrm{nn}}|k\rangle \sim k^{\alpha}$ applies with $\alpha \approx 0.4$.

In addition to degree–degree correlations, which characterize the topology of the network, we can study correlations between node strengths. The average nearest neighbour strengths are given by $s^N_{\mathrm{nn},i} = (1/k_i)\sum_{j\in\mathcal{N}(v_i)} s^N_j$ and $s^D_{\mathrm{nn},i} = (1/k_i)\sum_{j\in\mathcal{N}(v_i)} s^D_j$ which, when averaged over all nodes in the network with strength approximately equal to s, gives the average strength of nearest neighbours $\langle s^N_{\mathrm{nn}}|s^N\rangle$ and $\langle s^D_{\mathrm{nn}}|s^D\rangle$. Whereas the degrees of two adjacent nodes are strongly correlated, we find that the strengths of two adjacent nodes in most cases are not. Figure 7 shows that the s^D dependence of $\langle s^D_{\mathrm{nn}}|s^D\rangle \sim (s^D)^{\alpha^D}$ can be divided into two parts, where the independence observed for small s^D crosses over at $s_x \approx 10^4$ to a linear relationship. This linear region can be understood by studying the proportion of node strength that is contributed by a single link. It turns out that for very strong links with $w^D > 10^4$, which make up 4.4% of all links, the strength of both adjacent nodes is determined almost entirely by the weight of this single link such that $s_i \approx w_{ij} \approx s_j$ [23]. This explains the linear trend in strength–strength correlations. The plot for $\langle s^N_{\mathrm{nn}}|s^N\rangle$ suggests a qualitatively similar picture, where the linear trend naturally sets in earlier in terms of the absolute value of s^N.

The extent of clustering around a node i is quantified by the (unweighted) clustering coefficient $C_i = 2t_i/[k_i\,(k_i - 1)\,]$, where t_i denotes the number of triangles around node i [4]. Empirical networks have been found to have fairly high average clustering coefficients, which can be seen as manifestation of the presence of three-point correlations. Typically, one looks at the average clustering coefficient as a function of degree $\langle C|k\rangle$, known as the clustering spectrum, as shown in figure 8. Here $\langle C|k\rangle \sim k^{-1}$ as is commonly found in many empirical networks [24].

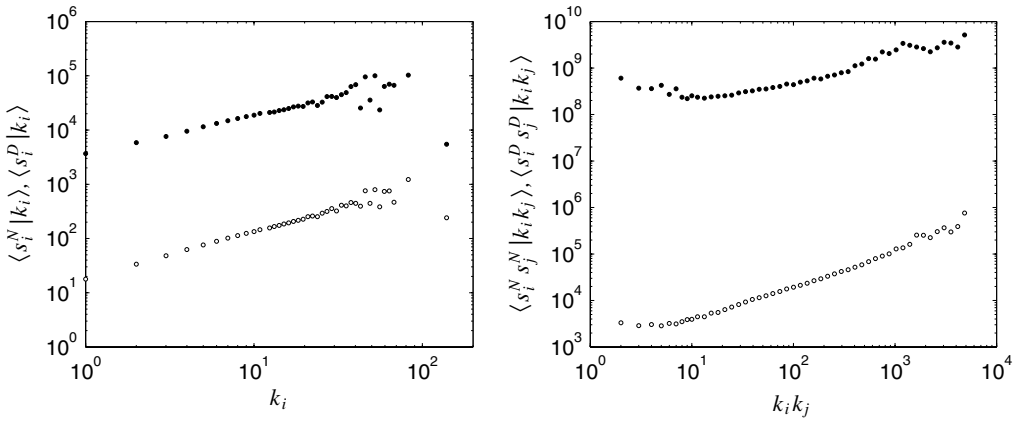

Figure 9. Average strength conditional on degree in terms of the number of calls $\langle s^N|k\rangle$ (○) and aggregated call duration $\langle s^D|k\rangle$ (●) (left) and average strength product s_is_j as a function of degree product k_ik_j denoted by $\langle s_i^N s_j^N|k_ik_j\rangle$ and $\langle s_i^D s_j^D|k_ik_j\rangle$.

This seems to indicate that clustering spectrum does not discriminate very well between different networks, which motivates us to adopt weighted network characteristics in section 3.

We have seen above that vertex degree distribution and vertex strength distribution are very similar in nature, which can be understood by examining degree-strength correlations. Average strength conditional on degree in terms of the number of calls $\langle s^N|k\rangle$ and aggregated call duration $\langle s^D|k\rangle$ are shown in figure 9. If there were no correlations between vertex degree and the weights of the links adjacent to the vertex, as can be obtained by shuffling the weights of the links, we would expect that $\langle s|k\rangle \sim k^\alpha$ with $\alpha = 1$, since $\langle s_i\rangle = k_i\langle w\rangle$, where $\langle w\rangle$ is the average link weight in the network. However, now we have $\langle s^D|k\rangle \sim k^{\alpha^D}$ where $\alpha^D \approx 0.8$ and $\langle s^N|k\rangle \sim k^{\alpha^N}$ where $\alpha^N \approx 0.9$, indicating that vertex strength grows somewhat more slowly than vertex degree, although α^N is close to one, i.e. linear growth. Based on the value of α^D, individuals who talk to a large number of friends appear to spend, on average, slightly less time per friend (lower w_{ij}^D) than those who spend less time on the phone.

We can study the strength product s_is_j as a function of degree product k_ik_j, the averages of which are denoted by $\langle s_i^N s_j^N|k_ik_j\rangle$ and $\langle s_i^D s_j^D|k_ik_j\rangle$, shown on the right in figure 9. In the absence of correlations, we would expect that $\langle s_is_j|k_ik_j\rangle = \langle w\rangle^2\langle k_ik_j\rangle$ giving $\langle s_j^D s_j^D|k_ik_j\rangle \sim (k_ik_j)^\beta$ with $\beta = 1$. However, we now obtain $\beta^D \approx 0.4$ whereas $\beta^N \approx 0.7$, corresponding to sublinear growth. Let us also introduce scaling exponents for degree products such that $\langle w_{ij}^N|k_ik_j\rangle \sim (k_ik_j)^{\gamma^N}$ and $\langle w_{ij}^D|k_ik_j\rangle \sim (k_ik_j)^{\gamma^D}$ and for strength products such that $\langle w_{ij}^N|s_i^N s_j^N\rangle \sim (s_i^N s_j^N)^{\delta^N}$ and $\langle w_{ij}^D|s_i^D s_j^D\rangle \sim (s_i^D s_j^D)^{\delta^D}$. The plots of these quantities are shown in figure 10. We find that $\gamma^D \approx -0.2$ and $\gamma^N \approx -0.1$, indicating that the links weights, whether measured in terms of w_{ij}^D or w_{ij}^N, are practically independent of the degree product k_ik_j. This shows that links weights are not determined by the absolute number of friends (node degrees) of v_i and v_j. In contrast, as we will see in section 4, link weights are dependent on the relative proportion of common neighbours (link overlap). For the latter exponents we have $\delta^N \approx \delta^D \approx 0.5$, such that w_{ij} scales as the geometric mean of the strengths of the adjacent nodes.

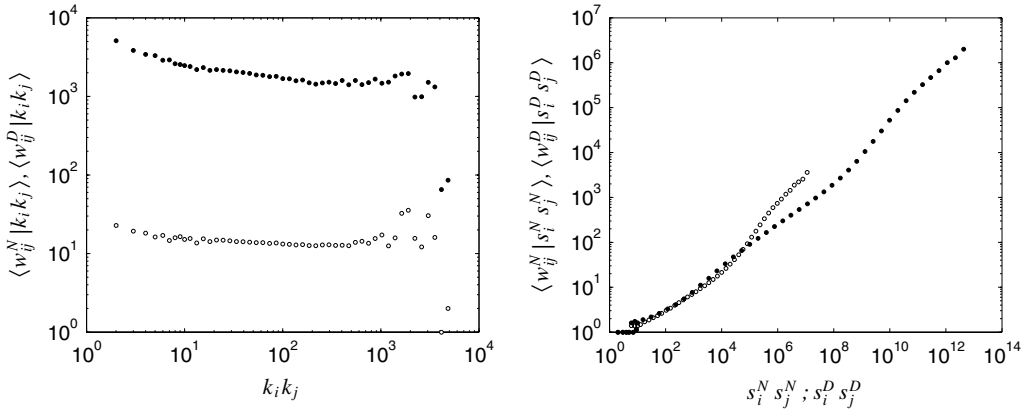

Figure 10. Scaling of weights w_{ij}^N (○) and w_{ij}^D (●) as a function degree product $k_i k_j$ (left) and strength product $s_i s_j$ (right).

Putting these structural properties together, we have seen that the network has a very steep degree distribution, resulting in few highly connected nodes, and even they are not as connected as hubs in scale-free networks are. The two weights, number of calls and aggregate call duration, are strongly correlated, and both yield steep strength distributions for nodes. This can be understood in light of the only slightly sublinear dependence of strength on degree, governed by the exponent α. Topologically the network is assortative, but not weight-assortative for a large majority of nodes. The weight of a given link is almost independent of the product of the degrees of adjacent nodes as governed by the almost vanishing exponent γ, but depends on the geometric mean of the strengths of the adjacent nodes as indicated by the value of exponent δ.

3. Advanced network characteristics

Study of purely topological properties of networks, as was done in section 2, is a useful starting point, but incorporating weights in the analysis is important, as it can enhance our understanding of the structural properties of the network. This motivates us to proceed to weighted network characteristics. Here important concepts are subgraph intensity and subgraph coherence that can be used to study the coupling between network structure and interaction strengths [25]. The intensity of subgraph g with vertices v_g and links ℓ_g is given by the geometric mean of its weights as

$$i(g) = \left(\prod_{(ij)\in\ell_g} w_{ij} \right)^{1/|\ell_g|}, \tag{1}$$

where $|\ell_g|$ is the number of links in ℓ_g [25]. Note that the unit of intensity is the same as the unit of network weights. To characterize the homogeneity of weights in a subgraph, we defined subgraph coherence $q(g)$ as the ratio of the geometric to the arithmetic mean of the weights as

$$q(g) = i(g)|\ell_g| / \sum_{(ij)\in\ell_g} w_{ij}. \tag{2}$$

Table 2. Number of cliques of order $k = 1, 2, \ldots, 10$ in the empirical network (empirical count) and their expectation values in a corresponding ER network (ER expectation) [28]. Note that $k = 1$ corresponds to the number of nodes $N = 6\,282\,226$ and $k = 2$ to the number of links $L = 16\,828\,910$, which are the same in the empirical and random network. These values of N and L give the link formation probability in the ER graph as $p = 2L[N(N-1)] \approx 8.5 \times 10^{-7}$. The expected number $E[X]$ of subgraphs with k nodes and ℓ links is given by $E[X] = \binom{N}{k}(k!/a)p^{\ell}$, where $\ell = k(k-1)/2$ and $a = k!$ is the number of graphs that are isomorphic to one another, i.e. automorphic, defined as adjacency-preserving permutation of the vertices of the graph [29]. Here, unlike elsewhere in the paper, the empirical network is a non-mutual one formed from the aggregated calls of 12 weeks. Note that subgraphs are counted multiple times, such that one subgraph of order k contains k subgraphs of order $k-1$ and so on. For example, one subgraph with $k = 5$ will also be counted as five instances of subgraph of order $k = 4$, and $5 \times 4 = 20$ instances of subgraph of order $k = 3$. The presence of high-order topological correlations, as manifest by the existence of cliques beyond order three (triangles) in the empirical network, makes is starkly different from an ER graph, in which high-order cliques have astronomically low probability to be present.

Order	Empirical count	ER expectation
1	6.3×10^6	6.3×10^6
2	17×10^6	17×10^6
3	5.6×10^6	2.6×10^1
4	1.4×10^6	2.5×10^{-11}
5	2.7×10^5	1.7×10^{-29}
6	4.5×10^4	7.8×10^{-54}
7	6.8×10^3	2.7×10^{-84}
8	799	7.0×10^{-121}
9	61	1.4×10^{-163}
10	2	2.0×10^{-212}

Here $q(g) \in [0, 1]$ and it is close to unity only if the weights of subgraph g do not differ much, i.e. are internally coherent [25].

The local structure of unweighted networks can be characterized by the appearance of small subgraphs, which have been related to the functionality of several networks [26, 27]. This is done by studying the number of times a subgraph of interest appears in the network, but to draw statistical conclusions about the appearance frequency of subgraphs, a reference system needs to be specified, which can be seen as analogous to setting up a null hypothesis H_0 in the statistics literature. The reference system is usually established by rewiring the network while conserving its degree distribution in order to remove local structural correlations present in the original network. Statistical significance of motifs is usually measured in terms of a z-score statistic [27]. Here we have chosen just to provide the number of fully connected subgraphs, i.e. k-cliques, up-to order $k = 10$ in table 2 for both the empirical network and a corresponding Erdős-Rényi (ER) network [28].

The motif framework has been generalized to weighted networks [25], with the motivation of studying the nature of coupling between interactions strengths (link weights w_{ij}) and local network topology (an ensemble of subgraphs g). We follow this line of enquiry and find intensity and coherence for some small fully connected cliques. As a point of comparison, we set up a *weight permuted reference* by simply shuffling the weights in the network, which removes weight correlations while leaving the network topology unaltered. Any deviation in motif intensities between the empirical and reference system has a straightforward interpretation: the local organization of weights in the empirical network is not random. While the z-score may be generalized to weighted networks as demonstrated in [25], it has the same shortcoming as the z-score has for unweighted networks, namely, that it is based on just one number characterizing the empirical network and two numbers characterizing the reference distribution. We follow an alternative approach here introduced in [30], which makes use of the entire intensity distribution $P^{\mathrm{E}}(i, g)$ for a subgraph g in the empirical network to the intensity distribution $P^{\mathrm{R}}(i, g)$ in the corresponding reference ensemble. Now the problem becomes one of comparing two distributions with one another for which several tools are available, such as the standard Kolmogorov–Smirnov test or the Kullback–Leibler divergence [31]. This approach suggests a shift in perspective from regarding subgraphs as discrete objects that either exist or not to a continuum of subgraph intensities and coherences.

Results are shown for intensity in figure 11 and for coherence in figure 12. Comparing the subgraph intensity distribution shows that the empirical subgraphs have considerably higher intensities than their random counterparts. Noting in particular the vertical logarithmic scale, we see that some high intensity subgraphs can be 10–1000 times more frequent in the empirical than in the reference ensemble. Especially for the larger subgraphs, e.g. $k = 6$, there are some extremely high intensity subgraphs in the empirical network, which are never created randomly in the reference ensemble. Similarly, the subgraphs in the empirical network are more coherent than their randomized counterparts. The differences become larger as we move to more complex subgraphs, the reason being that it is increasingly unlikely to create coherent subgraphs with many links by chance. Putting the results on intensity and coherence together, link weights within cliques are higher and more similar in magnitude than expected in a randomized reference. Consequently, there are important correlations between local network structure at the level of cliques, or communities, and interactions strengths within them.

Although the intensity-coherence framework is applicable to any arbitrary subgraph g, of which the k-cliques discussed above are an important class, perhaps one of the most important subgraphs in social networks are triangles. In the following we use the symbol Δ in place of g to emphasize that we are dealing with triangles, but point out that similar analyses could be carried out for any subgraph g. We denote the average intensity of triangles at node k by $\bar{i}_k(\Delta) = (1/t_k)\sum_{\Delta_k} i(\Delta)$, where $\sum_{\Delta_k}$ denotes a sum over all triangles containing node k. We can average this over all nodes that participate in one instance of a triangle, denoted by $\langle \bar{i}(\Delta)\rangle = (3n(\Delta))^{-1}\sum_k \bar{i}_k(\Delta)$, where $n(\Delta)$ is the number of triangles in the network and the factor three comes from the fact that a triangle consists of three nodes[7]. We emphasize that $\bar{i}_k(\Delta)$ denotes the mean intensity of triangles around a particular node k, where the mean is taken over all triangles attached to the node, whereas $\langle \bar{i}(\Delta)|s\rangle$ denotes average taken over all nodes whose strength is approximately s. The behaviour of average intensity of triangles as a function of node strength, $\langle \bar{i}(\Delta)^N|s^N\rangle$ and $\langle \bar{i}(\Delta)^D|s^D\rangle$, and average mean coherence, $\langle \bar{q}(\Delta)^N|s^N\rangle$

[7] In the case of an arbitrary subgraph g, the factor three would be replaced by $|v_g|$, the number of nodes in subgraph g.

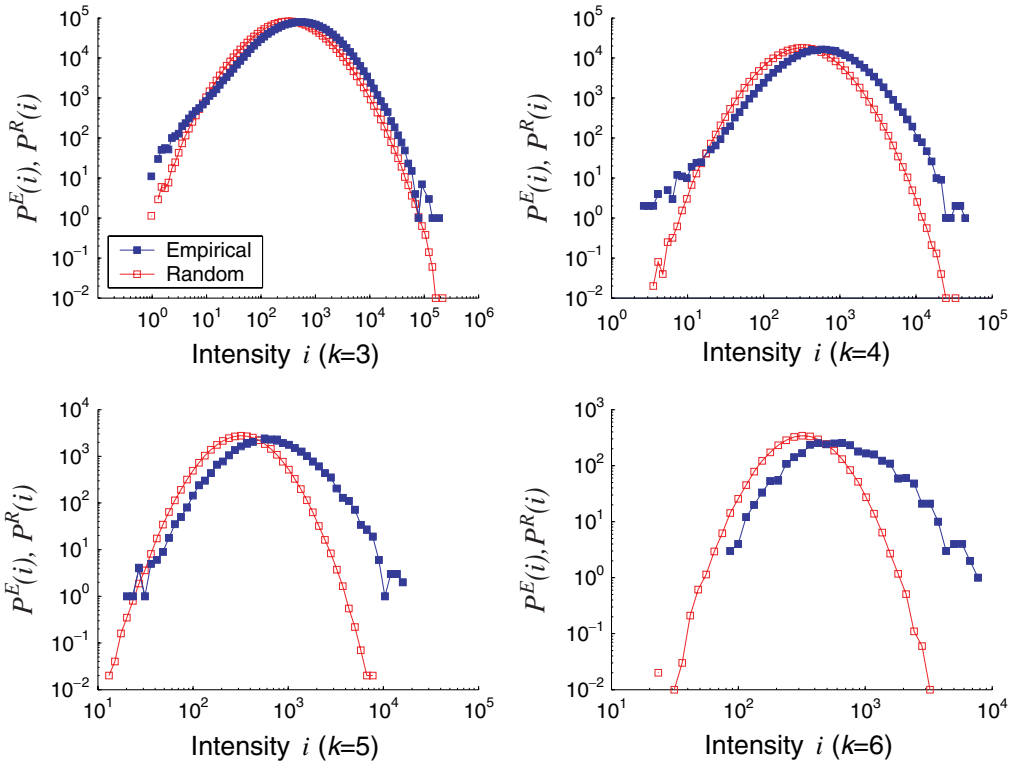

Figure 11. Distribution of subgraph intensity based on aggregate call duration weights w_{ij}^D for cliques g of order $k = 3, 4, 5, 6$ in the LCC of the empirical mutual network (solid blue squares) and in a reference ensemble (open red squares). Number of subgraphs of intensity i in the empirical network is given by $P^{\mathrm{E}}(i, g)$ and their average number in 100 realizations of the reference ensemble by $P^{\mathrm{R}}(i, g)$. A realization of the ensemble is obtained by shuffling the weights w_{ij}^D in the empirical network while keeping its topology fixed. All distributions are unnormalized and both horizontal and vertical scales vary between the panels.

and $\langle \bar{q}(\Delta)^D | s^D \rangle$, are shown in figure 13. We find that $\langle \bar{i}(\Delta)^N | s^N \rangle \sim (s^N)^{\epsilon^N}$, where $\epsilon^N \approx 0.5$ and $\langle \bar{i}(\Delta)^{D|s^D} \rangle \sim (s^{D)^{\epsilon^D}}$, where $\epsilon^D \approx 0.7$. The behaviour of average mean coherence $\langle \bar{q}(\Delta)^{D|s^D} \rangle$ is markedly different from that of the intensity, achieving a maximum at $s^{D \approx 10^3}$.

We can use the intensity of triangles to consider the effect of weights on the clustering properties of the network, and adopt the definition proposed for a weighted clustering coefficient in [25], leading to

$$\tilde{C}_i = \frac{1}{k_i\,(k_i - 1)} \sum_{j,k} \left(\hat{w}_{ij} \hat{w}_{ik} \hat{w}_{jk}\right)^{1/3} = C_i \bar{i}_i(\triangle), \tag{3}$$

where $\bar{i}_k(\triangle)$ again denotes the average intensity of triangles at node k. The weights are normalized by the maximum weight in the network, $\hat{w}_{ij} = w_{ij}/\max(w)$, required for reasons of compatibility with the topological clustering coefficient, and the contribution of each triangle depends on all of its edge weights [32, 33]. Note that the weighted clustering coefficient can be written as the

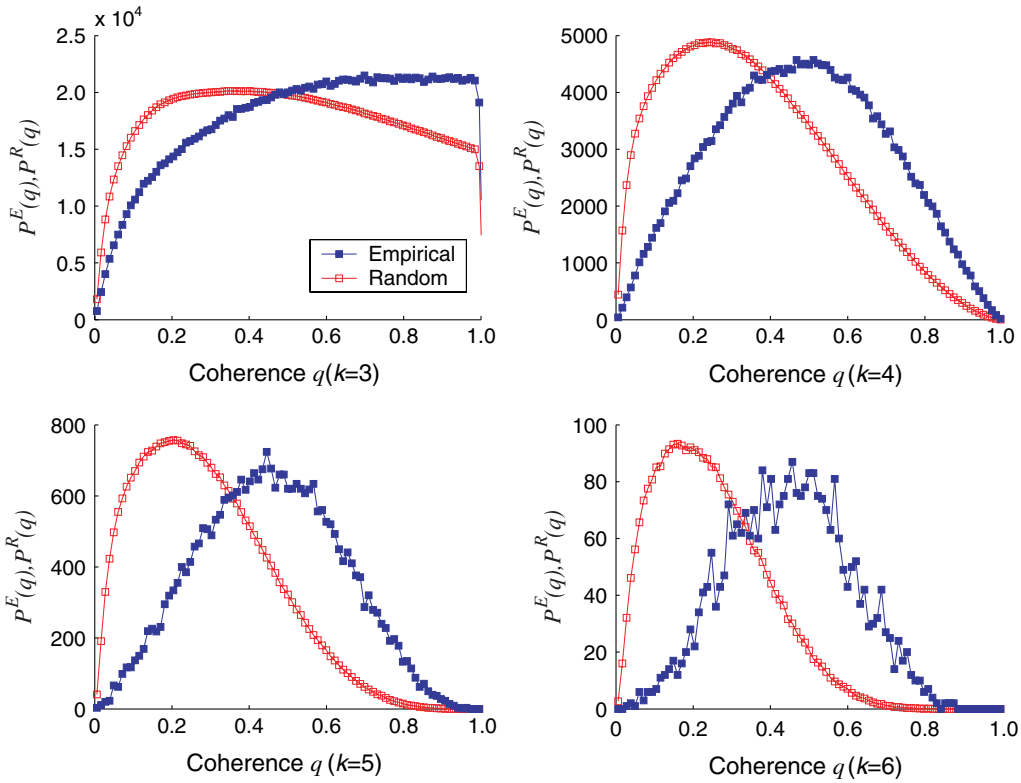

Figure 12. Distribution of subgraph coherence based on aggregate call duration weights w_{ij}^{D} for cliques of order $k = 3, 4, 5, 6$ in the LCC of the empirical mutual network $P^{\mathrm{E}}(q, g)$ (solid blue squares) and in the reference ensemble $P^{\mathrm{R}}(q, g)$ (open red squares). All distributions are unnormalized and both horizontal and vertical scales vary between the panels.

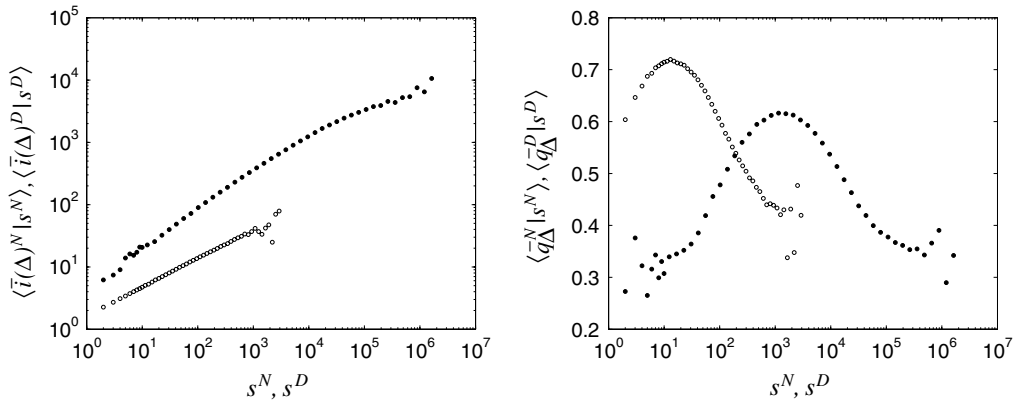

Figure 13. Average mean intensity of triangles $\langle \bar{i}(\Delta)^{N}|s^{N}\rangle$ and $\langle \bar{i}(\Delta)^{D}|s^{D}\rangle$(left), and average mean coherence $\langle \bar{q}(\Delta)^{N}|s^{N}\rangle$ and $\langle \bar{q}(\Delta)^{D}|s^{D}\rangle$ (right) as a function of node strengths s^{N} (○) and s^{D} (●).

product of the unweighted clustering coefficient and the average intensity of triangles at a node as shown in equation (3). Thus triangles in which each edge weight equals $\max(w)$ contribute unity to the sum, while a triangle having one link with a negligible weight will have a negligible contribution to the clustering coefficient. Results are shown in figure 8 next to the unweighted (topological) clustering coefficient. It is clear that the behaviour for number of calls and aggregate duration is very similar. For the duration we assume again that a crossover sets in at $s_x^D \approx 10^4$. Up-to this point the power law $\langle \tilde{C}|s^D\rangle \sim (s^D)^{\zeta^D}$ with $\zeta^D \approx 0.8$ gives an acceptable fit. However, the behaviour of $\langle \tilde{C}|s^N\rangle$ cannot really be described by a power-law.

4. Single link characteristics

Let us now move from subgraphs to study the properties of links and their immediate neighbourhood. We quantify the topological overlap of the neighbourhood of two connected nodes i and j by the relative overlap of their common neighbours, defined as

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}}, \tag{4}$$

where n_{ij} is the number of neighbours common to both nodes i and j [23]. It is worth pointing out that this is similar, but not identical, to the edge-clustering coefficient as introduced by Radicchi *et al* [34] as

$$C_{ij} = \frac{n_{ij}}{\min(k_i, k_j) - 1}, \tag{5}$$

where $\min(k_i, k_j) - 1$ is the maximum possible number of triangles around the (i, j) edge. Edge-clustering coefficient reflects the probability that a pair of connected vertices has a common neighbour, whereas overlap is the fraction of common neighbours a pair of connected vertices has. The reason for using O_{ij} as opposed to C_{ij} is that the denominator of equation (5) gives rise to two undesirable features in the context of social networks. Firstly, consider a subgraph in which vertices i and j are connected only with a single link such that $k_i = 1$ and $k_j > 1$, where vertex i is a leaf of the network. We now have $O_{ij} = 0$ indicating that these two individuals have no common friends, which seems a reasonable conclusion, whereas C_{ij} is either not defined or diverges as the denominator tends to zero. Secondly, consider a triangle (i, j, k) such that $k_i = 2$, $n_{ij} = 1$ and $k_j \geqslant 2$. If $k_j = 2$, then both $O_{ij} = 1$ and $C_{ij} = 1$. However, if $k_j > 2$, we still have $C_{ij} = 1$ for all values of k_j, whereas $O_{ij} = 1/(k_j - 1)$. This is to say that the overlap of common friends decreases as k_j increases since, although i and j still have just one common friend ($n_{ij} = 1$), the overlap of their common friends decreases as vertex j acquires new friends (k_j increases). This is again a reasonable feature of an overlap measure in a social context. Finally, as a general remark, since the overlap is a property of the link, it has the desirable property that, unlike C_{ij}, it is symmetric with respect to its arguments k_i and k_j.

The behaviour of average overlap as a function of absolute link weight $\langle O|w^D\rangle$ and cumulative link weight $\langle O|P_c(w)\rangle$ is shown in figure 14.[8] Average overlap $\langle O|w^D\rangle$ increases

[8] The cumulative link weight is defined in the following way. Let $P_<(x) = \int_{-\infty}^{x} p(w)\,dw$, where $p(w)$ is the probability density function for the link weights (either w^N or w^D). We define $P_c(w) = \psi \in [0, 1]$ if

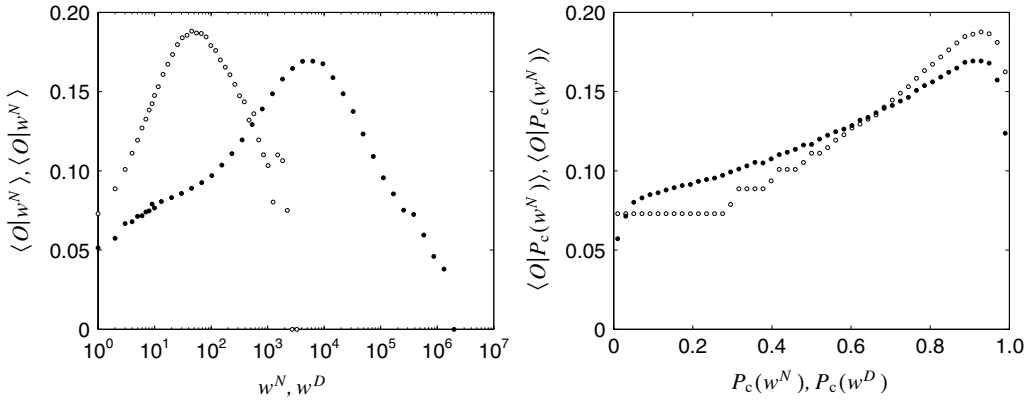

Figure 14. Average overlap as a function of absolute link weight $\langle O|w^D\rangle$ (left) and cumulative link weight $\langle O|P_c(w)\rangle$ (right) for w^N (○) and w^D (●).

up to $w^D \approx 10^4$, after which it declines strongly. However, $\langle O|P_c(w^D)\rangle$ shows that the declining trend is applicable to only some 5% of links, resulting from these individuals communicating predominantly just one other person as explored in [23]. Note that $s_x^D \approx 10^4$ was the crossover point in the distribution of $\langle s_{nn}^D|s^D\rangle$ and $\langle \tilde{C}|s^D\rangle$, indicating that the behaviour of these high-strength nodes is different from that of the rest. The high strength of these nodes derives from the top 5% of heavy links that also behave in an anomalous way as discussed in detail in [23].

Could the result concerning overlap O_{ij} versus link weight w_{ij} be affected by the fact that the phone call data is from a single operator and, consequently, calls to phone subscriptions managed by other operators are not included? Let us assume that an individual in the population has a probability $p = 0.2$ of having a subscription governed by the operator the data comes from. We assume that the nodes are all identical and that the probability of a node being governed by the operator is independent of the probability of its neighbour being governed by the operator. Given these assumptions, we can interpret p as the probability of a randomly chosen node being governed by the operator and, consequently, it being included in our network. Hence, the probability for a link to be included in the network is p^2 and that for a triangle is p^3. These probabilities give rise to expected number of nodes, links, and triangles in the unobserved population network as $\hat{N} = N/p = 5N, \hat{L} = L/p^2 = 25L$, and $\hat{T} = T/p^3 = 125T$, respectively. These numbers indicate that the expected number of links and triangles in the underlying network, to which we have only partial visibility by virtue of having a one-operator sample of it, are 25 times the number of links and 125 times the number of triangles in the observed network, respectively.

Since the value of p affects the number of observed nodes, links, and triangles in the sample, it is important to consider how it may affect overlap O_{ij}. To estimate the effect of p on $\langle O|w^D\rangle$, we follow an approach motivated by the Bootstrap-technique [35]. We generate a *resample* of the LCC of our network by including each node in the resample with probability p and by varying it obtain different sample sizes. In the limit of setting $p = 1$ we recover the original network. The results are shown in figure 15. Although lower values of p result in slightly lower values

$P_<^{-1}(\psi - \Delta\psi) < w \leqslant P_<^{-1}(\psi + \Delta\psi)$, where $P_<^{-1}(\cdot)$ is the inverse cumulative density function of link weights and $\Delta\psi = 1/50$. Note that $P_c(w)$ is different than $P_>(w)$, although we refer to both of them as cumulative link weight.

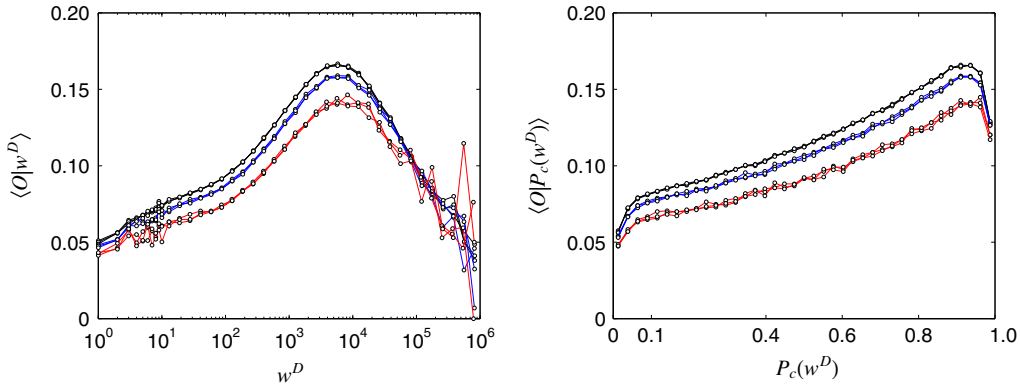

Figure 15. Average link overlap as a function of link weight $\langle O|w^D, p\rangle$ (left) and cumulative link weight $\langle O|P_c(w^D), p\rangle$ (right) for altogether nine network samples for the LCC of the mutual network. Three samples were drawn for each value of p, corresponding to the probability of a node in the initial network to be included in the sample. We used the values of $p = 0.8$ (top 3 curves), $p = 0.6$ (middle 3 curves), and $p = 0.4$ (bottom 3 curves), and the corresponding sample sizes were $N_{p=0.8} \approx 2.6 \times 10^6$, $N_{p=0.6} \approx 1.4 \times 10^6$, and $N_{p=0.4} \approx 0.4 \times 10^6$.

of $\langle O|w^D\rangle$, its qualitative behaviour is fairly insensitive to it. The cumulative plot shows how decreasing p does, in fact, cause the curve to become slightly flatter. This suggests that if the original network covered a larger fraction of the market or, alternatively, if data from several phone operators was aggregated, the value of $\langle O|P_c(w^D)\rangle$ would somewhat increase in absolute terms but, most importantly, its increasing trend with respect to w^D would become possibly even more pronounced. In short, the reported relationship between weight w and overlap O is not an artifact caused by having a sample from the underlying mobile phone call network.

A well-known hypothesis from sociology, the weak ties hypothesis of Granovetter, states that the proportional overlap of two individual's friendship networks varies directly with the strength of their tie to one another [36]. According to this hypothesis, the strength of a tie is a 'combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie' [36]. The present network is suitable for testing the weak tie hypothesis empirically at a societal level for two reasons. Firstly, the weights w_{ij}^D are phone call durations and thus implicate the time commitment to the relationship, one of the variables suggested to be indicative of the strength of an interpersonal tie. Secondly, the size of the network guarantees sufficient averaging and, therefore, produces reliable statistics. In addition, using the mutual network entails at least some degree of reciprocity (at least one call has been returned) and, importantly, commitment of phone time in this case also implies monetary costs to the caller. We find that the average overlap increases for about 95% of link weights, as shown in figure 14, and the behaviour of the remaining 5% can be accounted for (see supplementary information in [23]). Importantly, this increasing trend is practically unaffected whether number of calls w^N or aggregate call duration w^D are used as weights. Put together with the issue of sampling discussed above, these results provide a societal level verification of the weak ties hypothesis [23].

The results on overlap can be related to the concept of link betweenness centrality, defined for a link $e = (i, j)$ as $b_{ij} = \sum_{v \in V_s} \sum_{w \in V/\{v\}} \sigma_{vw}(e)/\sigma_{vw}$, where $\sigma_{vw}(e)$ is the number of shortest

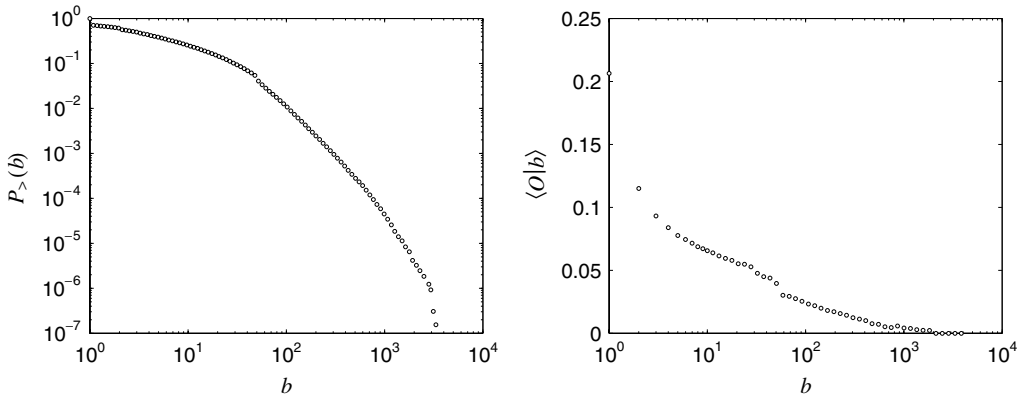

Figure 16. Cumulative distribution of link betweenness centrality $P_>(b)$ (left) in the LCC of the mutual network and the average link overlap as a function of link betweenness centrality $\langle O|b\rangle$ (right). Here $P_>(b)$ has been computed using a sample of $N_s = 10^5$ starting nodes from which the shortest paths to every other $N - 1$ nodes were found in order to calculate the betweenness centrality of links.

paths between v and w that contain e, and σ_{vw} is the total number of shortest paths between v and w [37]. In practice, we use the algorithm introduced in [38] to compute b_{ij} but, due to the heavy computational requirements of the algorithm, instead of using all the nodes of the set V making up the network, we use a sub-set of $N_s = 10^5$ nodes in the sample V_s as starting points. We then use the algorithm to find the shortest paths from these N_s nodes to all other remaining $N - 1$ nodes, every time keeping track of which links are used in constructing the shortest paths. Note that using this many source nodes results of the order of 10^{11} shortest paths to be computed in the network, more than a sufficient number, as was confirmed by using a smaller value for N_s. The cumulative distribution of link betweenness centrality is shown in figure 16. The figure also shows the behaviour of average link overlap as a function of link betweenness centrality $\langle O|b\rangle$. This is in full agreement with the above picture of the role of weak and strong links: weak links have low overlap but high betweenness centrality, reflecting their importance in holding the system together, while strong links have high overlap but low betweenness centrality and, as such, unlike the weak links, are not irreplaceable.

We give a summary of the studied basic, advanced, and single link network characteristics in table 3.

5. Percolation studies

We now turn to an examination of the implications of link removal on the global properties of networks, which has many precedents in the complex network literature [16, 37], [39]–[45]. However, instead of removing links randomly, we remove them based on either their weight w_{ij}, overlap O_{ij}, or betweenness centrality b_{ij} values. Removal can be carried out in one of two directions, i.e. either starting from links with low w_{ij}, O_{ij}, or b_{ij} values and proceeding towards higher ones or, alternatively, starting from links with high w_{ij}, O_{ij}, or b_{ij} and proceeding towards those with lower corresponding values. This thresholding process is governed by the control parameter f, the ratio of removed links, which allows us to interpolate between the initial

Table 3. Summary of basic, advanced, and single link characteristics used to study the LCC of the mutual network. The column titled 'scaling' gives, when appropriate, the approximate values obtained for the exponents found by fitting the the data to the given functional relation. The last column refers to the page where the characteristic is addressed.

Network characteristic	Notation	'Scaling'	See page
Degree distribution (cumulative)	$P_>(k)$	NA	6
Link weight distribution (cumulative)	$P_>(w^N)$, $P_>(w^D)$	NA	7,8
Node strength distribution (cumulative)	$P_>(s^N)$, $P_>(s^D)$	NA	7,8
Degree–degree correlation	$\langle k_{\rm nn}\|k\rangle$	$\sim k^\alpha$, $\alpha \approx 0.4$	8,9
Degree–degree correlation, w_{ij}^N–weighted	$\langle k_{\rm nn}^N\|k\rangle$		9
Degree–degree correlation, w_{ij}^D-weighted	$\langle k_{\rm nn}^D\|k\rangle$		9
Strength–strength correlation	$\langle s_{\rm nn}^N\|s^N\rangle$	$\sim (s^N)^{\alpha^N}$, see text	10
	$\langle s_{\rm nn}^D\|s^D\rangle$	$\sim (s^D)^{\alpha^D}$, see text	10
Clustering spectrum	$\langle C\|k\rangle$	$\sim k^{-1}$	10
Weight–degree correlation	$\langle s^N\|k\rangle$	$\sim k^{\alpha^N}$, $\alpha^N \approx 0.9$	10,11
	$\langle s^D\|k\rangle$	$\sim k^{\alpha^D}$, $\alpha^D \approx 0.8$	10,11
Strength product–degree product correlation	$\langle s_j^N s_j^N\|k_i k_j\rangle$	$\sim (k_i k_j)^{\beta^N}$, $\beta^N \approx 0.7$	11
	$\langle s_j^D s_j^D\|k_i k_j\rangle$	$\sim (k_i k_j)^{\beta^D}$, $\beta^D \approx 0.4$	11
Weight–degree product correlation	$\langle w_{ij}^N\|k_i k_j\rangle$	$\sim (k_i k_j)^{\gamma^N}$, $\gamma^D \approx -0.2$	11,12
	$\langle w_{ij}^D\|k_i k_j\rangle$	$\sim (k_i k_j)^{\gamma^D}$, $\gamma^N \approx -0.1$	11,12
Weight–strength product correlation	$\langle w_{ij}^N\|s_i^N s_j^N\rangle$	$\sim (s_i^N s_j^N)^{\delta^N}$, $\delta^N \approx 0.5$	11,12
	$\langle w_{ij}^D\|s_i^D s_j^D\rangle$	$\sim (s_i^D s_j^D)^{\delta^D}$, $\delta^D \approx 0.5$	11,12
Average triangle intensity–strength correlation	$\langle \bar{i}(\Delta)^N\|s^N\rangle$	$\sim (s^N)^{\epsilon^N}$, $\epsilon^N \approx 0.5$	15,16
	$\langle \bar{i}(\Delta)^D\|s^D\rangle$	$\sim (s^D)^{\epsilon^D}$, $\epsilon^D \approx 0.7$	15,16
Average triangle coherence–strength correlation	$\langle \bar{q}(\Delta)^N\|s^N\rangle$	NA	15,16
	$\langle \bar{q}(\Delta)^D\|s^D\rangle$	NA	15,16
Weighted clustering–strength correlation	$\langle \tilde{C}\|s^D\rangle$	$\sim (s^D)^{\zeta^D}$, $\zeta^D \approx 0.8$	10,16
	$\langle \tilde{C}\|s^N\rangle$	NA	10,16
Overlap–weight correlation	$\langle O\|w^N\rangle$	NA	16–18
	$\langle O\|w^D\rangle$	NA	16–18
Overlap–cumulative weight correlation	$\langle O\|P_{\rm c}(w^N)\rangle$	NA	16–18
	$\langle O\|P_{\rm c}(w^D)\rangle$	NA	16–18
Overlap–betweenness centrality correlation	$\langle O\|b\rangle$	NA	19,20

connected network ($f = 0$) and a set of isolated nodes ($f = 1$). We study the response of the network to removal of w_{ij}, O_{ij}, and b_{ij} links by monitoring four quantities as a function of the control parameter, which are (1) order parameter $R_{\rm LCC}$, the fraction of nodes in the LCC, (2) 'susceptibility' $\tilde{S} = \sum_s s^2 n_s / N$, where n_s is the number of clusters of size s, (3) average shortest path length $\langle \ell \rangle$, and (4) average clustering coefficient $\langle C \rangle$. Differences in the behaviour of these quantities reflect the global role different links have in the network.

The order parameter $R_{\rm LCC}$ is defined as the fraction of nodes in the LCC, i.e. the fraction of nodes that can all reach each other through connected paths. We find that removing links from low w_{ij} to high w_{ij} (red curve), from low O_{ij} to high O_{ij} (red curve), or from high b_{ij} to low b_{ij} (black curve) leads to a sudden disintegration of the network at $f^w = 0.8$, $f^O = 0.6$, and $f^b = 0.6$, respectively. In contrast, removing first the high weight, high overlap, or low

betweenness centrality links will shrink the network, but will not precipitously break it apart. This suggests that weak and strong links, low and high overlap links, and low and high betweenness centrality links have all different global structural roles in the network. In particular, it appears that removing low overlap links produces a qualitatively similar response to removing high betweenness centrality links.

The second row shows the behaviour of $\tilde{S} = \sum_s s^2 n_s / N$, which is analogous to magnetic susceptibility in thermal phase transitions, corresponding to the average component size in the network with the LCC excluded from the summation. According to percolation theory, if the network collapses via a phase transition at f_c, then $\tilde{S}$ diverges as $f \to f_c$ for an infinite system. A finite signature of such divergence is clearly visible in these plots upon removing low w_{ij}, low O_{ij}, or high b_{ij} links, suggesting that the network disintegrates at this point following a phase transition. Since the role of weak and strong ties is different at the local level and has important consequences from the sociological perspective [36], understanding their different global role is central, which is indeed a very pertinent question from the perspective of social network theory (see section 1). We have studied the global role of weak and strong links using finite size scaling (FSS) as reported in [23]. Although different FSS methods yielded slightly different results, removal of weak links (red curve) lead to a genuine phase transition at $f_c^w(\infty) = 0.80$, but there appears to be no phase transition when strong links are removed first (black curve). This result confirms that weak and strong links have qualitatively different global roles in social networks.

While the size of the largest component tells us about overall connectivity of the network, it does not convey information about its topology, only that the $N_{\mathrm{LCC}}(f = 0) R_{\mathrm{LCC}}(f)$ nodes are connected through one or more paths. One way to characterize the topology of the network is to study the average shortest path length, denoted by $\langle \ell \rangle$, which is the average number of links on the shortest path connecting any two vertices within the LCC. Note that as links are removed, the network becomes fragmented in components, of which we focus only on the largest one, i.e. the LCC for the given value of the control parameter f. Path lengths are also important from the perspective of network function and efficiency. The existence of a path between nodes is a necessary but not sufficient condition for there to be a flow of information between them. This is especially true if the transmission through links is leaky, i.e. it is possible for information to get lost along the way. Focusing on the role of weak and strong ties, we find that removal of weak ties increases path lengths more than removal of strong ties does, although the effect is stronger upon removing low O_{ij} or high b_{ij} links.

Path lengths are also related to the conjecture obtained from the weak ties hypothesis, according to which communities are locally connected by single weak ties, and removing these weak ties should therefore increase average path lengths making it more difficult to reach people [36]. Our result provides an empirical verification of the weak ties conjecture. It can also be related to a study dealing with search in social networks, according to which successful searches are conducted primarily through intermediate to weak strength ties without requiring highly connected hubs to succeed [8]. The present results suggest that the success of weak ties for search might lie in their function as community connectors, enabling one to reach outside of one's own community and thus expanding the set of individuals who may be reached through the network.

The average clustering coefficient $\langle C \rangle$ measures the local cliquishness of the network. Unlike the average shortest path length $\langle \ell \rangle$, which is computed only for the LCC for the given value of f, the average clustering coefficient is computed over all nodes in the network for which degree $k > 1$. Removing strong links (figure 17, row (g), black curve) leads to a convex

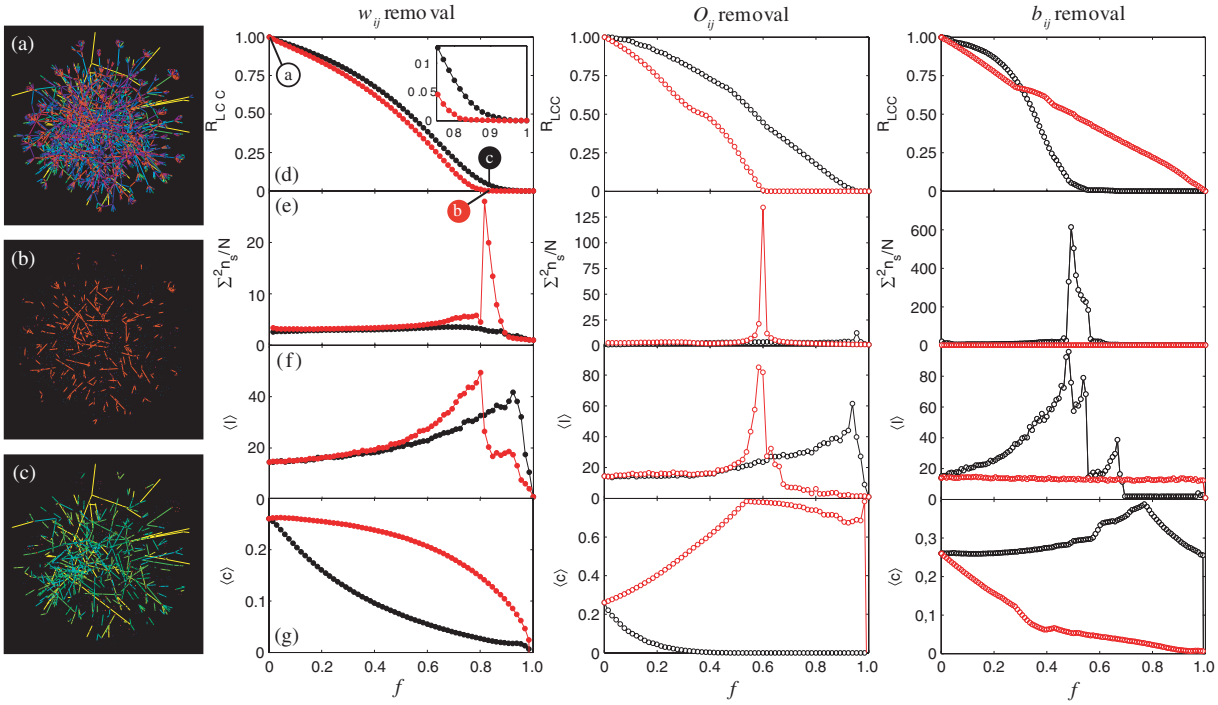

Figure 17. Percolation analysis. Panel (a) shows a small network sample with all links intact, (b) the same sample with 80% of the low w_{ij} links removed ($f = 0.80$, red curve), and (c) the sample with 80% of high w_{ij} links removed ($f = 0.80$, black curve). Rows (d)–(g): removal of high or low weight w_{ij} links (left column), overlap O_{ij} links (middle column), or betweenness centrality b_{ij} links (right column). The links are removed one at a time based on their ranking, such that the black curves correspond to starting removal from high w_{ij}, O_{ij}, and b_{ij} links, whereas the red curves represent the opposite, starting removal from low w_{ij}, O_{ij}, and b_{ij} links. The fraction of removed links is denoted by f. Row (d): the order parameter R_{LCC}, the fraction of nodes of nodes present in the LCC of the network for the given value of f to that present in the LCC for $f = 0$. Row (e): $\tilde{S} = \sum_s s^2 n_s/N$, corresponding to the average component size in the network with the LCC excluded from the summation. Row (f): average shortest path length $\langle \ell \rangle$ in the LCC of the system for the given value of f, which is also expected to diverge as $f \to f_c$. Row (g): average clustering coefficient $\langle C \rangle$ in the network.

clustering curve with an overall lower $\langle C \rangle$ than when weak links are removed. This happens because the strong links are mostly located in tightly connected communities where triangles are abundant. Consequently, removing them decreases the number of triangles and lowers clustering. Removing weak links (red curve) produces a concave clustering curve which first decreases very slowly. This is because the weak links are mostly located between communities, acting as local bridges and, therefore, rarely participate in triangles. Consequently, removing them has little effect on clustering. However, the difference in behaviour for overlap thresholding is even more drastic. On removing high O_{ij} links, the communities become shattered very quickly, so that at $f \approx 0.40$ average clustering coefficient is close to zero. The opposite happens on removing

low O_{ij} links. The average clustering increases up-to $f \approx 0.54$, compatibly with the fact that 53.5% of links in the LCC have $O_{ij} = 0$, and reaches a value almost as high as $\langle C \rangle \approx 0.80$. This results demonstrates quantitatively that the network is highly clustered and these clusters, or communities, can be filtered out reasonably well by removing low O_{ij} links. Again, removal of high overlap links is qualitatively similar to removing low betweenness centrality links and vice versa.

Since some community detection algorithms rely on the concept of betweenness centrality to detect communities [46], our results suggest that it may be possible to use the concept of overlap to detect communities at least in social networks. Bearing in mind that O_{ij} is a local characteristic and can be computed in $O(N)$, whereas b_{ij} is a global characteristic and takes $O(N^2 \ln N)$ to compute, algorithms relying on b_{ij} could use $1/O_{ij}$ as a local proxy for b_{ij}, potentially leading to significant gains in computing performance. We note that a modified version of the edge-clustering coefficient of equation (5) has also been used to replace edge betweenness centrality in a popular method for finding communities [34]. One could alternatively use O_{ij} without any modifications and, due to its desirable properties covered in section 4, it may be better suited for that purpose in identifying communities in social networks.

6. Discussion

Modern technologies enable the study of social networks of unprecedented size. A number of such investigations have appeared recently ranging from exploring email communication networks [6]–[8], [47] to identifying groups and strategies in an electronic marketplace [48]–[50]. In this paper, we constructed a network from mobile phone call records and used both aggregated call durations and the cumulative number of calls as a measure of the strength of a social tie. Since the network is derived exclusively from one-to-one communication, it can be used as a proxy for the underlying human communication network at the societal level which, to our knowledge, is the largest weighted social network studied as far.

In prototypical sociological studies the number of investigated individuals is limited to the order of a hundred [51], although exceptionally, like in the case of the Add Health database [52] as employed, for example, in [53], tens of thousands of individuals may be reached using questionnaires. This method enables a broad spectrum of interpersonal relations to be covered, although the subjectivity and quantification of interaction strengths are major problems. In this paper we have followed a complementary approach by basing the network on a specific type of social interaction, a phone call, allowing an objective measure of interactions for millions of people. We believe that studies like this one can provide valuable lessons about the large-scale structure of societies emerging from microscopic social interactions.

One of our focal points was to explore the relationship between local network topology and the associated weights. This is particularly important from the point of view of sociology, where such a relation was hypothesized a long time ago. In order to test the weak ties hypothesis, we used the concept of link overlap to measure the coupling between link weight and the overlap of the neighbourhood in the vicinity of the tie. We demonstrated that for 95% of the links the overlap and tie strength are correlated, verifying the hypothesis at a societal level. Moreover, we found the link overlap to be negative correlated with its betweenness centrality, suggesting that the former can be used as a local proxy for the latter, computationally heavy, global quantity.

We explored further the role of weights in the network using the concepts of intensity, coherence, and weighted clustering coefficient. We found correlations between local network structure at the level of cliques, or communities, and interaction strengths within them. The weighted clustering coefficient provides an appropriate tool for probing the strength of clustering due to weights, and may be used to differentiate between weighted networks that have fundamentally different coupling between network topology and interaction strengths. We found that the network is assortative in terms of topology as expected but, rather surprisingly, is not weight-assortative for a large majority of nodes. Further, the coupling between local network structure and interaction strengths carries over to the global level. We quantified this by studying the differences in percolation behaviour depending on the properties of the removed links. Following this approach we also verified the so-called weak ties conjecture, a global manifestation of the weak ties hypothesis.

The obtained results can be used as a basis for devising weighted models of social networks. In particular, the relation between topological and statistical properties should be incorporated in such models. This enables collective social phenomena to be studied, such as spreading of information and opinion formation, at a level of realism and scale not possible in the past. The lessons learnt from this endeavour are not limited to understanding human societies, but may find application in other domains as well. Finally, we believe that our systematic approach can be adopted to study other weighted networks, and the present results can be seen as a reference against which other networks may be compared.

Acknowledgments

JPO is grateful to the ComMIT Graduate School, and Finnish Academy of Science and Letters, Väisälä Foundation, for a travel grant to visit A-L Barabási at Harvard University. This research was partially supported by the Academy of Finland, Centers of Excellence programs, project no. 44897 and 213470 and OTKA K60456. GS and ALB were supported by the NSF-ITR, NSF-DDAAS projects and by the McDonell Foundation.

References

[1] White H C, Boorman S A and Breiger R R 1976 Social structure from multiple networks. I. Blockmodels of roles and positions *Am. J. Sociol.* **81** 730–780

[2] Granovetter M 1992 *Decision Making: Alternatives to Rational Choice Models Economic Action and Social Structure: The Problem of Embeddedness* (Newbury Park, CA: SAGE) pp 304–33

[3] Wasserman S and Faust K 1994 *Social network analysis: methods and applications Structural Analysis in the Social Sciences* (Cambridge: Cambridge University Press)

[4] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2

[5] Albert R, Jeong H and Barabási A-L 1999 The diameter of the world wide web *Nature* **401** 130

[6] Ebel H, Mielsch L-I and Bornholdt S 2002 Scale-free topology of e-mail networks *Phys. Rev.* E **66** 35103(R)

[7] Eckmann J-P, Moses E and Sergi D 2004 Entropy of dialogues creates coherent structures in e-mail traffic *Proc. Natl Acad. Sci. USA* **101** 14333–7

[8] Dodds P S, Muhamad R and Watts D J 2003 An experimental study of search in global social networks *Science* **301** 827–9

[9] Aiello W, Chung F and Lu L 2000 A random graph model for massive graphs *Proc. 32nd ACM Symp. on the Theory Comput.* (New York: ACM) pp 171–80

[10] Watts D J 2007 A twenty-first century science *Nature* **445** 489

[11] Haythornthwaite C 2005 Social networks and Internet connectivity effects *Inform. Commun. Soc.* **8** 125–47
[12] Lee S H, Kim P-J and Jeong H 2006 Statistical properties of sampled networks *Phys. Rev.* E **73** 16102
[13] Dunbar R I M 1992 Neocortex size as constraint on group size in primates *J. Human Evolution* **22** 469–93
[14] Dorogovtsev S N and Mendes J F F 2003 *Evolution of Networks* (Oxford: Oxford University Press)
[15] Pastor-Satorras R and Vespignani A 2001 Epidemic spreading in scale-free networks *Phys. Rev. Lett.* **86** 3200–3
[16] Cohen R, Erez K, ben Avraham D and Havlin S 2000 Resilience of the internet to random breakdowns *Phys. Rev. Lett.* **85** 4626–8
[17] Colizza V, Barrat A and Barthélemy M Vespignani A 2006 The role of airline transportation network in the prediction and predictability of global epidemics *Proc. Natl Acad. Sci USA* **103** 2015–20
[18] Almaas E, Kovács B, Vicsek T, Oltvai Z N and Barabási A-L 2004 Global organization of metabolic fluxed in the bacterium escherichia coli *Nature* **427** 839–43
[19] Goh K-I, Kahng B and Kim D 2001 Universal behavior of load distribution in scale-free networks *Phys. Rev. Lett.* **87** 278701
[20] Pastor-Satorras R, Vázquez A and Vespignani A 2001 Dynamical and correlation properties of the internet *Phys. Rev. Lett.* **87** 258701
[21] Newman M E J 2002 Assortative mixing in networks *Phys. Rev. Lett.* **89** 208701
[22] Barrat A, Barthélemy M, Pastor-Satorras R and Vespignani A 2004 The architecture of complex weighted networks *Proc. Natl Acad. Sci. USA* **101** 3747–52
[23] Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J and Barabási A-L 2007 Structure and tie strengths in mobile communication networks *Proc. Natl Acad. Sci. USA* **104** 7332–6
[24] Szabó G, Alava M and Kertész J 2004 Complex networks *Clustering in Complex Networks* (*Springer Lecture Notes in Physics* vol 650) ed E Ben-Nain *et al* (Berlin: Springer) pp**104** 139–62
[25] Onnela J-P, Saramäki J, Kertész J and Kaski K 2005 Intensity and coherence of motifs in weighted complex networks *Phys. Rev.* E **71** 065103
[26] Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Sheffer M, Ayzenshtat I and Alon U 2004 Superfamilies of designed and evolved networks *Science* **303** 1538–42
[27] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 Network motifs: simple building blocks of complex networks *Science* **298** 824–7
[28] Erdős P and Rényi A 1960 *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17
[29] Bollobás B 2001 *Random Graphs* 2nd edn (Cambridge: Cambridge University Press)
[30] Onnela J-P 2006 Complex networks in the study of financial and social systems *PhD thesis* Helsinki University of Technology
[31] Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86
[32] Saramäki J, Onnela J-P, Kertész J and Kaski K 2005 Characterizing motifs in weighted complex networks *Science of Complex Networks, AIP Conf. Proc* ed J F F Mendes *et al* p 108
[33] Saramäki J, Kivelä M, Onnela J-P, Kaski K and Kertész J 2007 Generalizations of the clustering coefficient to weighted complex networks *Phys. Rev.* E **75** 027105
[34] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D 2004 Defining and identifying communities in networks *Proc. Natl Acad. Sci. USA* **101** 2658–63
[35] Efron B and Tibshirani R J 1994 *An Introduction to the Bootstrap* (London: Chapman and Hall)
[36] Granovetter M 1973 The strength of weak ties *Am. J. Sociol.* **78** 1360–80
[37] Holme P 2002 Edge overload breakdown in evolving networks *Phys. Rev.* E **66** 36119(R)
[38] Newman M E J 2001 Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality *Phys. Rev.* E **64** 16132(R)
[39] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A and Wiener J 2000 Graph structure in the web *Comput. Netw.* **33** 309–20
[40] Albert R, Jeong H and Barabási A-L 2000 Error and attack tolerance of complex networks *Nature* **406** 378–82
[41] Callaway D S, Newman M E J, Strogatz S H and Watts D J 2000 Network robustness and fragility: percolation on random graphs *Phys. Rev. Lett.* **85** 5468–71

[42] Cohen R, Erez K, ben Avraham D and Havlin S 2001 Breakdown of the internet under intentional attack *Phys. Rev. Lett.* **86** 3682–85
[43] Watts D J 2002 A simple model of global cascades on random networks *Proc. Natl Acad. Sci. USA* **99** 5766–71
[44] Holme P and Kim B J 2002 Vertex overload breakdown in evolving networks *Phys. Rev.* E **65** 066109(R)
[45] Onnela J-P, Kaski K and Kertész J 2004 Clustering and information in correlation based financial networks *Eur. Phys. J.* B **38** 353
[46] Girvan M and Newman M E J 2002 Community structure in social and biological networks *Proc. Natl Acad. Sci. USA* **99** 7821–26
[47] Wilkinson D, Tyler J and Huberman B 2003 Email as spectroscopy: Automated discovery of community structure within organizations *Communities and Technologies* ed M Huysman, E Wegner and V Wulf (Dordrecht: Kluwer)
[48] Yang I and Kahng B 2006 Bidding process in online auctions and winning strategy: rate equation approach *Phys. Rev.* E **73** 67101
[49] Yang I, Oh E and Kahng B 2006 Network analysis of online bidding activity *Phys. Rev.* E **74** 016121
[50] Reichardt J and Bornholdt S 2005 Ebay users form stable groups of common interest *Preprint* physics/0503138
[51] Giddens A 2006 *Sociology* 5th edn (Cambridge: Polity Press)
[52] The National Longitudinal Study of Adolescent Health online at http://www.cpc.unc.edu/addhealth
[53] Gonzalez M C, Herrmann H J, Kertesz J and Vicsek T 2007 Community structure and ethnic preferences in school friendship networks *Preprint* physics/0611268