Li, Ming Xia; Palchykov, Vasyl; Jiang, Zhi Qiang; Kaski, Kimmo; Kertész, János; Miccichè, Salvatore; Tumminello, Michele; Zhou, Wei Xing; N Mantegna, Rosario

Statistically validated mobile communication networks

Published in:
New Journal of Physics

DOI:
10.1088/1367-2630/16/8/083038

Published: 01/01/2014

Document Version
Publisher's PDF, also known as Version of record

PAPER • OPEN ACCESS

Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data

New Journal of Physics
The open access journal at the forefront of physics
Deutsche Physikalische Gesellschaft DPG | IOP Institute of Physics

Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data

Ming-Xia Li[1], Vasyl Palchykov[2,3,4], Zhi-Qiang Jiang[1], Kimmo Kaski[2], János Kertész[2,5], Salvatore Micciché[6], Michele Tumminello[7], Wei-Xing Zhou[1] and Rosario N Mantegna[5,6,8]

[1] School of Business, School of Science and Research Center for Econophysics, East China University of Science and Technology, Shanghai 200237, People's Republic of China
[2] Department of Biomedical Engineering and Computational Science, Aalto University, FI-00076 Aalto, Finland
[3] Institute for Condensed Matter Physics, National Academy of Sciences of Ukraine, UA-79011 Lviv, Ukraine
[4] Instituut-Lorentz, Universiteit Leiden, 2300 RA Leiden, The Netherlands
[5] Center for Network Science, Central European University, Nador 9, H-1051, Budapest, Hungary
[6] Dipartimento di Fisica e Chimica, Università di Palermo, Viale delle Scienze, Ed. 18, I-90128, Palermo, Italy
[7] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Viale delle Scienze, Edificio 13, I-90128 Palermo, Italy
[8] Department of Economics, Central European University, Nador 9, H-1051, Budapest, Hungary
E-mail: rn.mantegna@gmail.com

Received 22 March 2014, revised 14 June 2014
Accepted for publication 24 June 2014
Published 21 August 2014

New Journal of Physics **16** (2014) 083038
doi:10.1088/1367-2630/16/8/083038

Abstract

Big data open up unprecedented opportunities for investigating complex systems, including society. In particular, communication data serve as major sources for computational social sciences, but they have to be cleaned and filtered as they may contain spurious information due to recording errors as well as interactions, like commercial and marketing activities, not directly related to the social network. The network constructed from communication data can only be considered as a proxy for the network of social relationships. Here we apply a systematic method, based on multiple-hypothesis testing, to statistically validate

the links and then construct the corresponding Bonferroni network, generalized to the directed case. We study two large datasets of mobile phone records, one from Europe and the other from China. For both datasets we compare the raw data networks with the corresponding Bonferroni networks and point out significant differences in the structures and in the basic network measures. We show evidence that the Bonferroni network provides a better proxy for the network of social interactions than the original one. Using the filtered networks, we investigated the statistics and temporal evolution of small directed 3-motifs and concluded that closed communication triads have a formation time scale, which is quite fast and typically intraday. We also find that open communication triads preferentially evolve into other open triads with a higher fraction of reciprocated calls. These stylized facts were observed for both datasets.

Keywords: complex networks, social systems, statistically validated networks, mobile call records, 3-motifs

1. Introduction

The data deluge has its origin in the development of information communication technology, which in turn has revolutionized the scientific research into social systems. The 'digital footprints' that we leave behind in almost all of our activities enable unprecedented investigations, in both depth and sample size. A new discipline, called 'computational social science' [1], has emerged to join the efforts of social scientists, computer scientists, physicists, and mathematicians in a truly interdisciplinary approach with the aim of better understanding the laws of human society both at an individual and at a collective level.

Mobile call records (MCRs) play a special role in the studies of human societies, as the mobile phone coverage is close to 100% in the adult population in much of the world and mobile phones are our companions in almost all of our activities. Accordingly, MCRs are well suited for mapping out the structure of social networks [2, 3], including the dynamics [4] and hierarchical structure [5] of the communities, for studying dynamic aspects of human behavior like mobility characteristics [6–8] or communication patterns [9, 10], and temporal motifs [11]. The origin of mobile call records shows an increasing socio-economic and cultural variety, ranging across different European [2, 5], American [7], Asian [12] and African [13] sources. Although no systematic comparison has yet been made, some universal features seem to emerge. These include the Granovetterian structure [2, 14] of the network, the bursty character of communication [15], and the strong inhomogeneity both in the number of contacts and in the strength of the activities [16].

MCRs do indeed provide detailed information about human interactions: data for millions of users as regards who called whom, when, how long the conversation took and the whereabouts of the callers serve as a goldmine for understanding the structure of the society and the dynamic laws of communication and mobility of individuals. Moreover, if so-called metadata are at the disposal of the researchers, deeper insight into gender and age related behavioral patterns can be mapped out [17, 18].

Mesoscopic structures like network motifs [19] are of particular interest for the understanding of the structure and function of the society. A motif is a set of isomorphic

subgraphs and it is generally assumed that it represents a functionally important group of nodes if its cardinality significantly exceeds the expected number of such subgraphs in a reference system, which is usually the configuration model. With this concept, it was possible to identify classes of networks, where within one class (e.g., networks of genetic transcription in different species or networks of different languages) there is similar over-representation of motifs. On the basis of weight intensity, the concept of motifs has also been generalized to weighted networks [20]. Recently, there has been a growing interest in the dynamic patterns. Dynamic motifs [11, 18] are classes of similar event sequences, where the similarity refers not only to the topology but also to the temporal order of the events (e.g., phone calls). Using the information about the locality of the calls, mobility motifs were defined and the classification of human mobility patterns was enabled [21].

In temporal networks [25], where links are present only temporarily for an interaction event, static motifs defined on the aggregate networks present a time evolution as a function of the aggregation window. So far this has not been investigated, although this dynamics contains interesting information about the system. Time stamped mobile phone data are particularly suitable for such a study. One then asks: what are the typical motifs that are over-expressed in the network of MCRs? How do they emerge as a function of time? What is the characteristic time needed for the evolution of the motifs? These are the questions that we will focus on in this paper.

In social science there is a long tradition for studying triads, i.e. subgraphs of three nodes connected by directed links [22]. Recently, triads have been investigated in many other classes of complex networks ranging from biological to economic and financial networks [23, 24]. Following the terminology introduced by Alon and collaborators, triads are called 3-motifs. Motifs of higher order (typically of order 4 or 5) are also sometime considered. However, the most developed theories and the largest number of empirical studies about motifs concern 3-motifs. In fact the number of different motifs explodes when motifs of higher order are considered and the investigation of 4-motifs or 5-motifs in large networks is extremely demanding from a computational point of view. In social sciences a large amount of studies have focused on the 3-motif statistics and dynamics with the aim of using this information to detect global properties of the social system investigated. So far, the most prominent property investigated in social networks is triadic closure. The triadic closure is observed when a triad with only two relationships detected among the three social actors evolves to another triad with all the pairwise relationships present to some degree [22].

In the present study, we investigate the 3-motifs observed in two large databases of MCRs. Specifically, we investigate MCRs of two different mobile companies, one operating in Europe and the other in China. In this way we are studying the effectiveness of our approach and the validity of our investigation of communication links of social origin in two datasets that are different in various respects, e.g. as regards the telecom company (with its specific commercial policy), the geographical location and the recording time period. We have chosen to focus our investigation on the directed 3-motifs for two main reasons: (i) because they are extremely informative from a social point of view and (ii) because a reliable empirical estimation requires a series of strict conditions in the processing of very large samples.

A major problem with big data is that they have to be cleaned and filtered, as they contain spurious information due to recording errors and interactions, like commercial and marketing activities, not directly related to the study at hand. Usually MCRs are not collected for scientific purposes and, even if the companies attempt to provide the relevant data, there could be serious

problems. One example is that for studying social relationships, private communication is needed; however, experience tells us that sometimes phones registered as private are used for professional purposes, like in call centers or marketing and information campaigns. In fact, the presence of large spurious communication hubs, e.g. large call centers, significantly alters the statistics of 3-motifs (and, more generally, of any class of motifs). Dialing wrong numbers is another possible source of false links. In addition, the usual corruption arising during coding, transferring and processing data can also take place. Unless data are cleaned, spurious links could be misinterpreted as real social relationships. This problem is part of the general topic of information filtering in complex networks with strong inhomogeneities [26]. In fact, human related systems usually show properties changing over many orders of magnitude, and this is so for communication networks also: the distributions of degrees or activities are fat tailed [16].

A somewhat arbitrary way of filtering data was introduced by Onnela *et al* [2]. Three measures were taken: (i) only mutual connections were considered as links, i.e., both individuals had to initialize calls during the period of observation; (ii) links with total call duration of less than 10 s during the period of 18 weeks examined were ignored; and (iii) the nodes with less than 60 s of total call activity were filtered out. In fact, in this way spurious nodes with enormous (10^4) numbers of unidirectional connections and sometimes more than 24 hours/day (!) activity were eliminated. The 10 s cutoff served to filter out the calls of wrong numbers. However, this method unintentionally distorts the results, as there can be many socially relevant unidirectional links and even short duration links that may carry social interaction.

Another, more systematic way of filtering was proposed by Serrano *et al* [27]. The idea is to statistically validate the links by deciding locally which of the links carry a disproportionate fraction of the weights adjacent to a given node. Comparing the empirical observations with a null model that takes into account the inhomogeneities of the system can reveal significant over-representation of links, thus indicating their relevance. Carrying out the procedure node by node results in what is called the 'multiscale backbone' of the system. This method indicates important aspects of filtering, including the necessity of statistical validation and the relevance of the selection of an appropriate null model. However, it is asymmetric for the nodes of the links and it has some restrictions upon the degree k of the nodes (isolated links between two $k = 1$ nodes can never be validated, irrespective of the weight of the link), and it handles the local network topology independently of the rest of the network.

The problem of pruning and/or filtering of a network has also been encountered in the construction and analysis of similarity based networks. In fact, the construction of similarity based networks can be seen as a filtering procedure selecting informative structures of the underlying system, such as minimum spanning trees [28], partial correlation interdependence [29], subgraphs of arbitrary genus [30], planar graphs [31], etc.

Recently, a method for filtering out statistically links in bipartite complex networks [33] was proposed. As the mobile call network can be considered as a bipartite one, where one set of nodes corresponds to the mobile phone users and the other one to the calls that they perform, the method can be straightforwardly applied to our problem. As it is based on multiple-hypothesis testing, global information is built in, and thus the above-mentioned problems can be avoided. This method has already been applied successfully to a number of systems, including the networks of simple organisms, financial stocks and the Internet Movie Database [33], classification of investor strategies [34] and of the specializations of criminal suspects [35].

In this paper, we adapt and apply the method introduced in [33] to mobile phone communication networks obtained from MCRs of two different regions, which are a European country and the province level municipality of Shanghai (China). We first construct the communication networks from the raw MCRs (for short, called by us 'original networks') and then the networks of the statistically validated links, also called Bonferroni networks. We keep the directed character of the links, as it carries important information about the underlying social relationships. The comparison between the original and the Bonferroni networks shows significant differences in the basic statistical properties. For example, our filtering removes the extremely large hubs (which would contradict the social brain hypothesis [36]) but keeps a large number of unidirectional contacts.

The Bonferroni filtering of the original network allows us to perform a detailed analysis of so-called 3-motifs. We show that the study of 3-motif statistics and dynamics is unreliable unless we perform the Bonferroni filtering on the original network. This is due to the fact that the empirical estimation of 3-motifs is strongly affected by the presence of huge communication hubs that have no social origin but only some socio-technical motivation, such as in the case for call centers. We study in the Bonferroni network the time evolution of the communication 3-motifs. Our results show that communication 3-motifs are typically characterized by triadic closure at an intraday time scale. In fact, 3-motifs with only links detected between two pairs of subscribers primarily evolve into other 3-motifs characterized by a higher number of reciprocated calls. Triadic closure is preferentially observed in communication 3-motifs only after the calls of the open 3-motif are reciprocated.

The paper is structured as follows. In the next section, we discuss the application of the Bonferroni network method to the MCRs. In section 3, we describe our results for the directed 3-motifs. Then the temporal evolution of the motifs is discussed. Finally we present the conclusions.

2. The Bonferroni network of mobile call records

In order to analyze communication data for constructing MCR based networks, one has to first decide whether the entries in the records serve as good proxies for real social interactions in a probabilistic sense. This is a multiple-hypothesis test validation problem, which we approach by adapting and applying a directional version [32] of the recently introduced method of Bonferroni networks [33].

2.1. Data

We investigate two sets of data: one from a Chinese mobile phone service provider and another one from a European service provider. The Chinese data contain time stamped data of all (hashed) subscribers of the service within the time periods from 28 June to 24 July 2010, and from 1 October to 31 December 2010. In the second period, the calls recorded on 12 October, 5, 6, 13, 21, and 27 November, and 6, 8, 21, and 22 December contain missing records, and these days are removed from our analysis. Thus we have in total 109 days of calls recorded for the Chinese data. This dataset consists of 4031 090 subscribers and 1091 695 590 calls (done with both subscribers and non-subscribers of the service provider). When we select calls occurring only among subscribers, the number of calls reduces to 128 410 897, i.e., 88.24% of the calls go

to non-subscribers. The set of mobile phone users including subscribers and non-subscribers exceeds nine million users.

The data from the European provider contain all records of its 7387 034 subscribers during 212 days between 1 January 2007 and 31 July 2007. This includes 3969 043 426 calls, 682 124 009 of which occurred between subscribers of the given provider, i.e., 81.54% of all calls connecting subscribers with non-subscribers. The whole set of subscribers and non-subscribers exceeds 91 million users.

As the primary focus of our investigations is on the evolution of 3-motifs, in the present study we perform our investigations on the calling networks of subscribers only. In fact, including non-subscribers would alter the 3-motif statistics because calling data between two non-subscribers are not recorded in our datasets.

2.2. Statistically validated networks

In our directed network, nodes are mobile phone subscribers and a directional link is set from subscriber A to subscriber B if A makes a call to B in a selected time window, the weight w of the link being the number of calls in the time window investigated. For each link in the network, we perform a statistical test to check whether a link is statistically validated against a null hypothesis assuming heterogeneous calling profiles of the subscribers. The method is a directional variant of the method introduced in [33]. The statistical test is implemented as follows. We define N as the total number of calls among the subscribers in the system, and focus on two subscribers, i and j, to check whether the numbers of calls of i to j are over-expressed with respect to a null hypothesis taking into account heterogeneity in the number of calls performed. Let us call n_c^i the number of calls made by subscriber i as a caller, and n_r^j the number of calls that subscriber j receives. By labeling the number of times subscriber i calls j with X, the probability of observing X calls between them is given by

$$H\left(X\middle|N, n_c^i, n_r^j\right) = \frac{\binom{n_c^i}{X}\binom{N-n_c^i}{n_r^j-X}}{\binom{N}{n_r^j}}.$$

We can therefore associate a p-value with the observed number $X = n_{cr}^{ij}$ of calls from subscriber i to subscriber j as follows:

$$p\left(n_{cr}^{ij}\right) = \sum_{X=n_{cr}^{ij}}^{\min\left[n_c^i, n_r^j\right]} H\left(X\middle|N, n_c^i, n_r^j\right).$$

Calculating the p-value for all the directed edges, which are N_E in our network, implies that we run N_E statistical tests for obtaining the network. When a large number of statistical tests are performed simultaneously the effectiveness of the statistical test can be decreased by a large number of false positives unless a multiple-hypothesis test correction is used. In the present study, we use the Bonferroni correction, which is the strictest multiple-hypothesis test correction controlling the familywise error rate when either dependent or independent multiple hypotheses are tested. That means that the univariate level of statistical significance $p_u = 0.01$ must be replaced by a multivariate level, to be set as $p_m = p_u/N_E = 0.01/N_E$.

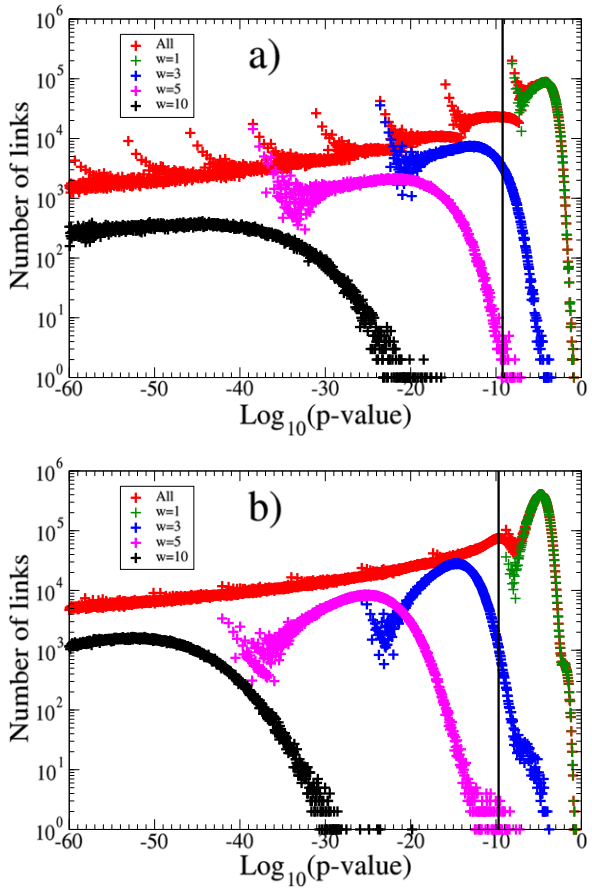

Figure 1. Number of links as a function of the p-value for the Chinese (panel (a)) and European (panel (b)) datasets. The red symbols describe the histogram for all links. Symbols of different color refer to the number of links of pairs of callers and receivers with weight equal to 1 (green), 3 (blue), 5 (purple) and 10 (black). The vertical line indicates the Bonferroni threshold. Links located to the left of the threshold are retained in the Bonferroni network. The network is obtained by considering the entire period. Only links between subscribers are considered.

If the estimated $p(n_{cr}^{ij})$ is less than p_m, we conclude that the link from subscriber i calling subscriber j is not due to the high heterogeneity of the subscribers and most probably reflects a social interaction between the subscribers. Accordingly, we set a link from i to j in the filtered network that is named the Bonferroni network.

We show in figure 1 series of histograms of the number of links characterized by a certain p-value for the Chinese and the European datasets respectively. Different histograms (characterized by different colors) are obtained by grouping the various links in terms of the number of calls characterizing them, i.e., in terms of their weights. The time interval used to build the original network is the entire time period available, that is 109 days and 212 days for the Chinese and European datasets respectively. Figure 1 shows that in both datasets the links with just one call (weight equal to 1) are characterized by a p-value which is larger than the Bonferroni threshold (indicated as a vertical line). The links that are filtered out from the original network comprise essentially all the links with weight 1 and some of the links with weight up to 5. Under the conditions of our analysis, both for the Chinese and for the European

datasets, when the weight is larger than 5 the links are always included in the Bonferroni network (see the case of $w = 10$ in figure 1).

The fact that links with unit weight are not present in the Bonferroni network is due to the procedure of statistical validation of the links. In fact, for weight 1, the above defined p-value reads

$$p_1 = \sum_{X=1}^{\min[n_c^i, n_r^j]} H\left(X \middle| N, n_c^i, n_r^j\right) \tag{1}$$

$$= 1 - H\left(0 \middle| N, n_c^i, n_r^j\right) \geqslant 1 - H(0|N, 1, 1) = H(1|N, 1, 1) \tag{2}$$

$$= \frac{1}{N} \geqslant \frac{0.01}{N_E} = p_m \Leftrightarrow 100 \geqslant \frac{N}{N_E}, \tag{3}$$

and the last inequality holds true in our system, because the average number of calls per directed link, $\frac{N}{N_E}$, is much smaller than 100. As mentioned above, with our statistical validation procedure also some of the links with higher weights do not get validated in the Bonferroni network. However, the absence of validation is not a direct consequence of the small average number of phone calls per link. For example, let us consider a simple case in which $n_r^j = n_c^i = n_{cr}^{ij} = 2$. In this case, the p-value is

$$p_2 = \sum_{X=2}^{2} H(X|N, 2, 2) = H(2|N, 2, 2) = \frac{2}{N(N-1)}. \tag{4}$$

This p-value would be statistically significant if it were $<0.01/N_E$, and such a condition is easily attained, even in a sparse system like the present one:

$$p_2 = \frac{2}{N(N-1)} < \frac{0.01}{N_E} \Leftrightarrow \frac{200}{N-1} < \frac{N}{N_E}. \tag{5}$$

Indeed, the latter inequality says that, to validate the link from i to j, it is sufficient if just the average number of calls per link is larger than $\frac{200}{N-1}$, which is a quantity smaller than 1 in any setting that includes more than 201 phone calls.

In the setting of the Bonferroni threshold there is a margin of arbitrariness. In other words, which is the most appropriate threshold to be used when we obtain distinct daily networks that we wish to compare? We believe that the answer to this question depends on the type of comparison that one aims to perform on the networks obtained. Therefore there might be more and less restrictive choices in the setting of the Bonferroni threshold. To minimize the number of false positive links, in the present study we set the Bonferroni threshold to $p_m = 0.01/N_E$, where N_E is the number of edges observed in the overall periods investigated (we have a single period investigated for the European data[9] and two distinct periods investigated in the Chinese data[10]). In

[9] For the European dataset the Bonferroni threshold is set to 0.01/49 029 577.

[10] For the Chinese dataset, we have two periods of data. In the first period (the first month) 6441 490 links are present in the original network between 2309 619 subscribers and in the second period (the last three months) 13 616 634 links are present between 3492 116 subscribers. We consider the two time periods as separate time periods and we set two Bonferroni thresholds as 0.01/6441 490 and 0.01/13 616 634 when performing the construction of the Bonferroni networks. For the original daily networks, we have about 6.53×10^5 nodes and 6.94×10^5 links on average. By using the statistical test, 52.17% of the nodes and 65.87% of the links are removed in daily networks on average.

this way, the Bonferroni threshold is rather conservative for the networks computed at short time intervals, e.g., at daily and weekly time intervals.

It should be noted that the choice of the Bonferroni threshold within broad limits up to some orders of magnitude does not crucially affect the composition of the networks obtained for the different time intervals. For instance, if the Bonferroni threshold used to construct daily Bonferroni networks for the Chinese data is increased by one order of magnitude, then the number of statistically validated links increases by only 4.8% on average.

2.3. Basic metrics and the degree distribution

We show in figure 2 the time evolution of some basic network indicators for the daily networks obtained from MCRs of the European dataset. The top panel shows the number of nodes (subscribers) which are present in the original (red line) and in the Bonferroni (blue line) network. Roughly half of the nodes present in the original network are also present in the Bonferroni network. A weekly pattern is clearly seen in both curves and some special days are also observed (see for example day 5 and day 105). The middle top panel shows the number of links. For the set investigated and for the time period considered, the Bonferroni network retains roughly one third of the links of the original network. The middle bottom panel shows the percentage of nodes present in the largest connected component. On average the original network has a largest connected component including 12% of the nodes. A large weekly cycle of amplitude close to 5% is observed. It is worth noting that the largest connected component of the Bonferroni network has a negligible fraction of the nodes. This means that the Bonferroni network shows a large number of disconnected clusters of subscribers without a giant component. The interconnection is provided by the presence of large hubs and weak ties that are filtered out in our approach.

This behavior is specific to the daily networks. The percentage of nodes in the largest connected component increases when the period of time used to detect the network increases. We show in table 1 the average percentage and the standard deviation of the nodes which are present in the largest connected component of the original and Bonferroni networks obtained for different time periods for the Chinese and European datasets. From the table we see that for weekly networks the percentage of nodes of the largest connected component is already almost 49% and almost 35% for the Chinese and European datasets respectively. These values further increase for the monthly networks when the largest connected components of the Bonferroni networks are 73% and 81%, i.e., values not too different from the ones (81% and 94%) observed in the original networks for the Chinese and European datasets respectively. We interpret these results as an indication that our filtering methodology is able to detect a progressively increasing fraction of the weak ties that provide the interconnections building the largest connected component. This observation is in agreement with the detected role of weak ties discussed in [2].

The bottom panel of figure 2 shows the time evolution of the number of 3-motifs. In the original network this number is fluctuating and presents a huge spike at day 79. In the case of the Bonferroni network, the time evolution is fluctuating less and no spike is present. In summary, our statistical validation procedure selects a network characterized by properties that are much more stable than those of the original network. We hypothesize that the Bonferroni network is able to retain links whose social motivations are typically more pronounced than those of the ones left out from the original network; thus the Bonferroni network is a better

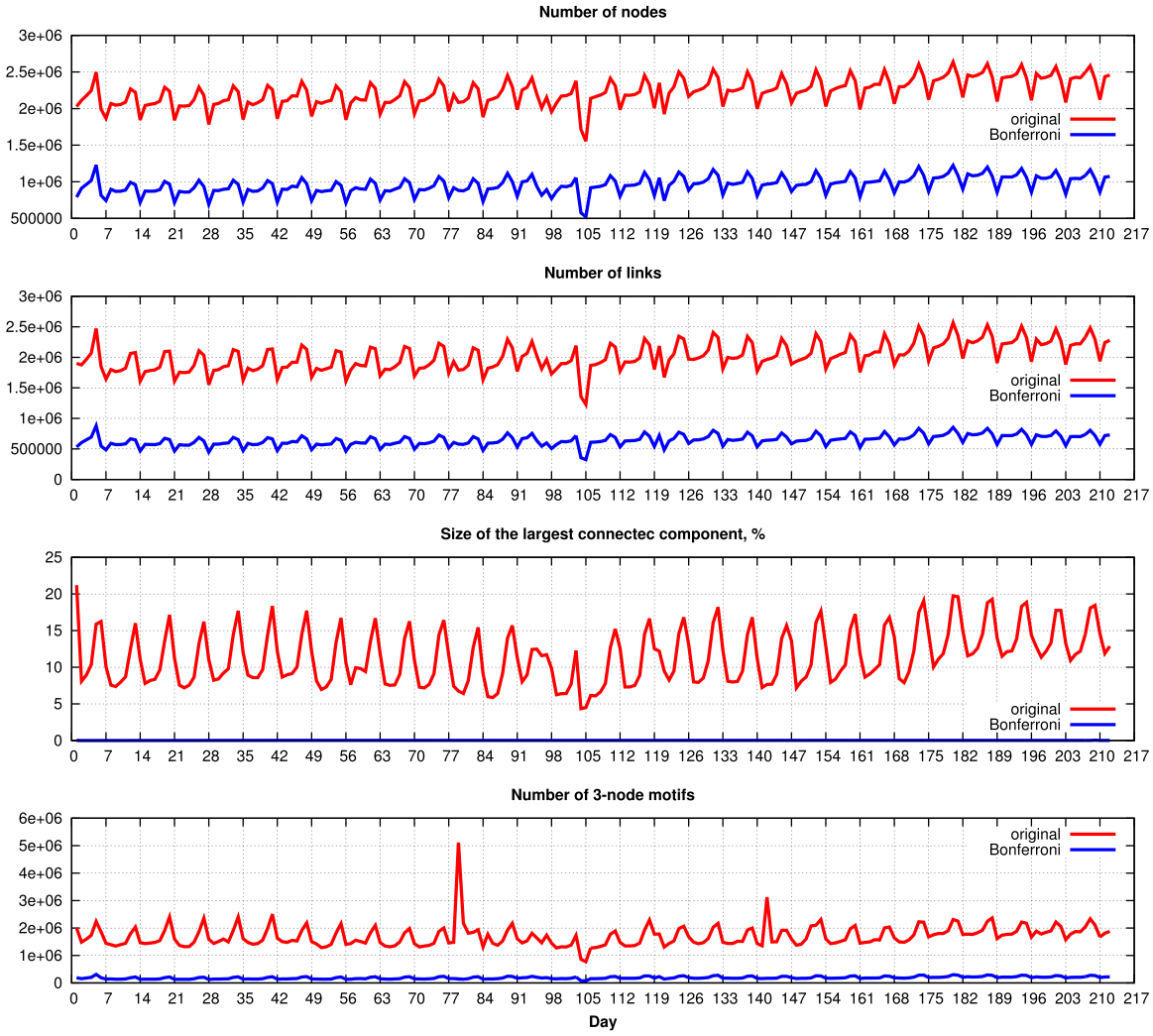

Figure 2. Time evolution of some basic network indicators for the daily networks obtained from MCRs of the European dataset. The top panel shows the number of nodes (subscribers) which are present in the original (red line) and in the Bonferroni (blue line) network. A weekly pattern is clearly seen in both curves and some special days are also observed (see for example day 5 and day 105). The middle top panel shows the number of links. The daily Bonferroni networks retain roughly one third of the links. The middle bottom panel shows the percentage of nodes present in the largest connected component. The original network has a largest connected component including on average 12% of the nodes. A large weekly cycle is observed. The largest connected component of the Bonferroni network has a negligible fraction of the nodes. The bottom panel shows the time evolution of the number of 3-motifs. In the original network this number is fluctuating and presents a huge spike at day 79. In the case of the Bonferroni network, the time evolution is fluctuating less and the huge spike is not present.

proxy for the underlying social network than the original one. Our hypothesis is supported by the results that we obtain for some important network metrics and for the census of the 3-motifs and their dynamics.

Table 1. Summary of the average values of the percentages of nodes present in the largest connected components of daily, weekly, monthly and complete networks both for the networks of the original set and for Bonferroni networks. We also report the standard deviation (as a percentage) of the average value. The results were obtained for the Chinese (top) and European (bottom) datasets. Subscribers only.

	Original				Bonferroni			
	Daily	Weekly	Monthly	All	Daily	Weekly	Monthly	All
Mean (%)	29.33	68.62	81.47	85.69	0.41	48.96	72.50	79.12
SD (%)	8.53	2.66	2.55	—	0.26	4.82	2.47	
Mean (%)	11.45	75.85	93.89	98.85	0.018	34.53	81.35	96.79
SD (%)	3.69	1.77	0.44	—	0.008	4.79	2.38	—

We show in figure 3 the cumulative in-degree and out-degree distributions for the Chinese datasets (subscribers only) for the entire period (109 days). The cumulative distributions are shown for the original network (panel (a)) and for the Bonferroni network (panel (b)). We observe a series of interesting differences in the cumulative distributions between the original and the Bonferroni networks. We observe in the original network subscribers with very large out-degree (of the order of 3000) and in-degree (of the order 700). We also note that the tails of the in-degree and out-degree distributions are very pronounced and quite different from each other (with the out-degree distribution significantly more pronounced that the in-degree distribution). In the case of the Bonferroni network the in-degree and out-degree distributions are still showing pronounced tails, but the largest degree is of the order of 200. We also note that the tails of the in-degree and out-degree distributions are in the Bonferroni case similar, with the in-degree being only slightly more pronounced than the out-degree for very large degrees.

A similar pattern is observed also in the degree distributions of European data (see panels (c) and (d) of figure 3). However, we also note differences between the European and the Chinese distributions. Specifically, for the original network (panel (c) of figure 3) the more pronounced tail is observed for the in-degree distribution in the European case, whereas the opposite is observed in the Chinese case. We also note that in the European case the distributions of the Bonferroni networks show slightly different tails: the tail of the in-degree distribution is more pronounced than that of the out-degree case. It is worth noting that, similarly to what we observe for the Chinese dataset, the maximal in-degree is close to 250 and the maximal out-degree is close to 150 if we do not consider an in-degree outlier characterized by a degree of 1131. The tails of the cumulative distributions of the in-degree and out-degree in the European case are well described by a power law decay with an exponent equal to 3.85 and 6.25 respectively[11].

3. 3-motifs

We show in figure 4 the 13 different kinds of 3-motifs that can be observed in a network. There are different ways to code the identities of these motifs. In the present paper, we use the labeling

[11] The exponent for the out-degree distribution is pretty big. It decays so fast that it is difficult to distinguish between power law and exponential decay.

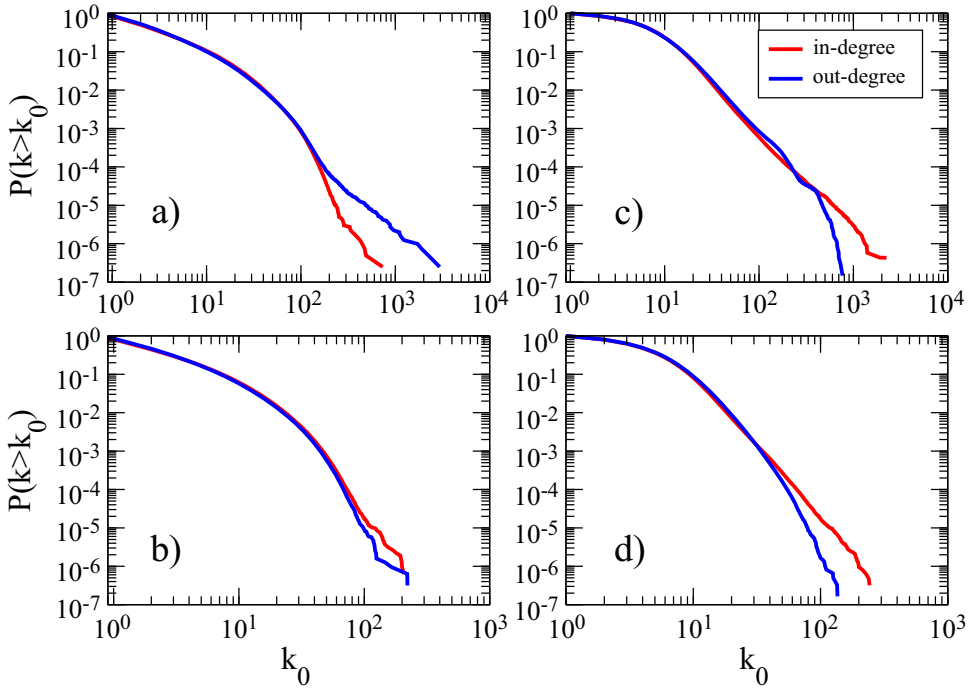

Figure 3. Cumulative in-degree (red line) and out-degree (blue line) distributions. In the left panels, we show results for the Chinese dataset, entire period (109 days), subscribers only. Panel (a): original network. Panel (b): Bonferroni network. In the right panels, we show results for the European dataset, entire period (212 days), subscribers only. Panel (c): original network. Panel (d): Bonferroni network. A single occurrence with in-degree equal to 23348 is not shown in panel (c) and a single occurrence with in-degree equal to 1131 is not shown in panel (d).

Figure 4. List of directed 3-motifs.

of Milo *et al* [19]. Here, we investigate the properties of communication 3-motifs of networks obtained from MCRs data.

A direct inspection of figures 5 and 6 shows that the estimation of the fraction of daily 3-motifs for the original network presents seasonalities of various frequencies and huge spikes localized at specific weeks. The seasonality is extremely pronounced for 3-motifs presenting only two of the three possible relationships (see the panels of figure 5). On the other hand, the pattern observed in the Bonferroni networks is more stable and shows only a weekly seasonality and a small deviation occurring for some special days (days with labels 5 and 105, which are most probably related to big holidays). In the Bonferroni network, the weekly pattern is quite evident for the 3-motifs with two pair relationships (figure 5), whereas for triads with a triangle structure the weekly pattern is less evident, especially in some cases—for example, for the 3-motif labeled as 98 (see figure 6).

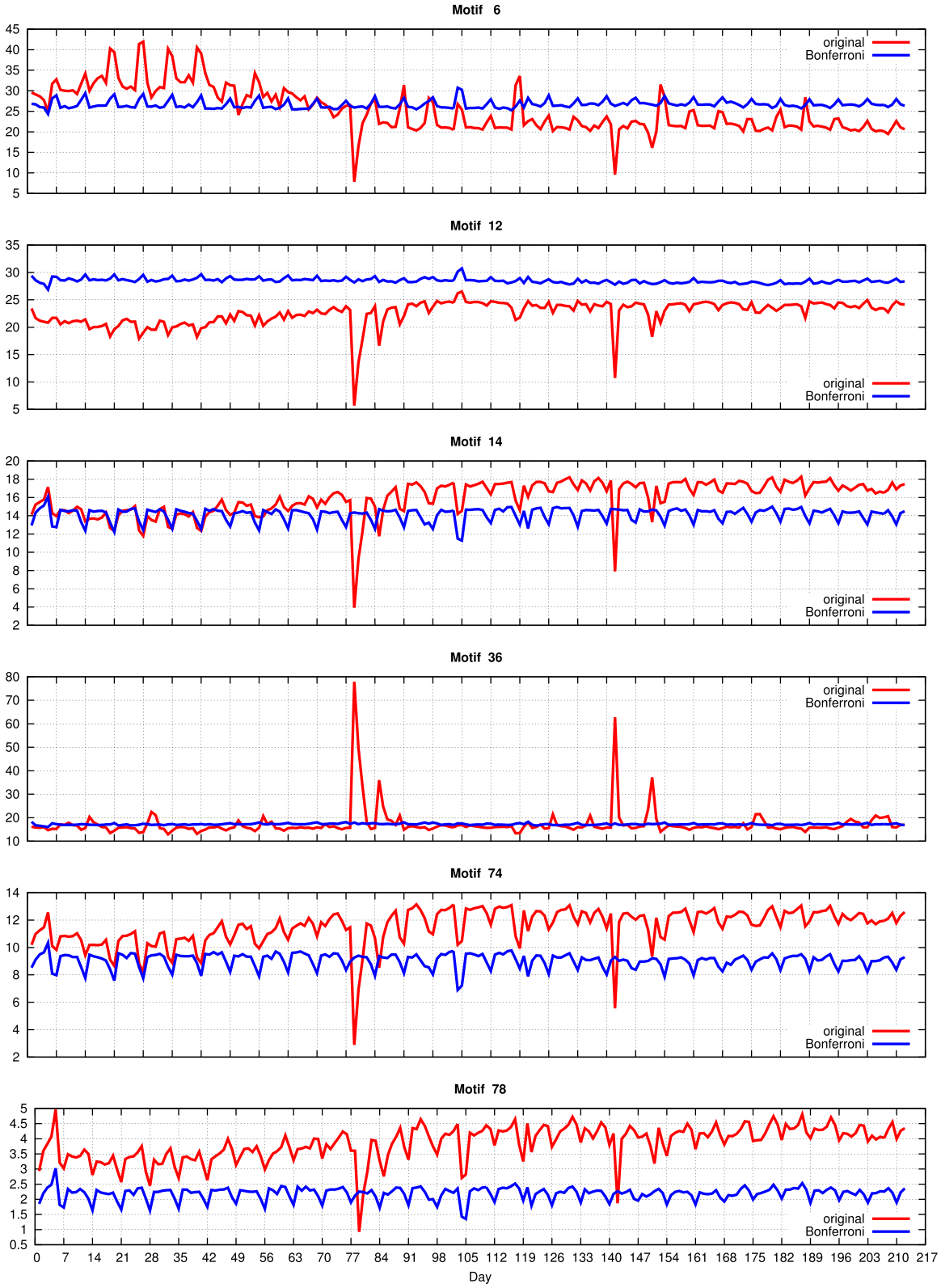

Figure 5. Fraction of 3-motifs observed in each day of the European data set. From top to bottom we have 3-motifs encoded as 6, 12, 14, 36, 74 and 78 respectively. The red line refers to the original network whereas the blue line refers to the Bonferroni network.

We interpret these empirical results as supporting our hypothesis that the Bonferroni network is sampling relationships characterized by a strong social interaction, whereas the original network also includes kinds of calls that are related to commercial or technical activities such as the ones typically performed by call centers. The presence of these activities can substantially alter counts of the triads because a node with a very large in-degree or out-degree participates in a large number of triads. This kind of spurious effect is clearly not observed in the Bonferroni network.

We present in tables 2 and 3 summary statistics for the fractions of 3-motifs observed in the daily, weekly and monthly original and Bonferroni networks for the Chinese and European datasets respectively. For each 3-motif and for each network we report the average value observed for real data μ and the average value $\mu_{\rm rnd}$ observed by randomly shuffling the network a large number of times while keeping the in-degree and out-degree of each node fixed and keeping the number of bidirectional relationships constant. The counting of the 3-motifs and the shuffling procedures were performed by using the FANMOD algorithm [37]. We also report in the tables the standard deviation observed for real data σ and for shuffled data $\sigma_{\rm rnd}$.

For each 3-motif we evaluate a z-score defined as $z = (\mu - \mu_{\rm rnd})/\sigma$. This variable indicates the deviation of the observed average value from the average value obtained by random shuffling of the network in units of the standard deviation. We have decided to use this definition of the z-score instead of the another possible one, namely $z_2 = (\mu - \mu_{\rm rnd})/\sigma_{\rm rnd}$, because our definition is the most conservative one in the present case. We highlight, in the tables, the average percentage of a 3-motif in boldface when its associated z-score is larger than 3 (a character **(+)** follows the average percentage value in this case) or smaller than -3 (a character **(−)** follows the average percentage value in this case). Tables 2 and 3 show that 3-motifs split into two sets. The first one is the set of 3-motifs showing communication arcs only between two of the three pairs of nodes of the motif, i.e. 3-motifs encoded as 6, 12, 14, 36, 74, and 78. The second set is the set of 3-motifs with all pairs showing at least one communication link (3-motifs encoded as 38, 46, 98, 102, 108, 110 and 238). Tables 2 and 3 show that for the first set of 3-motifs, the average percentage of the 3-motifs is close to the value expected for random connections (6, 12, and 36) or less than expected for the 3-motifs 14, 74, and 78. On the other hand, for the second set of motifs (38, 46, 98, 102, 108, 110 and 238), all the 3-motifs are presenting an average percentage which is higher than the value expected for random driven communications. In other words, the underlying social structure and the communication style of the social actors over-express the 3-motifs characterized by triadic closure. This behavior is observed at daily, weekly and monthly time scales (with a pattern more pronounced when the time period used to build the network is longer) and it is observed both for the Chinese and the European datasets.

The above cited results are qualitatively observed both for the original and for the Bonferroni networks. However, original and Bonferroni networks present values of the average percentages of the 3-motifs which are quite different, especially for weekly and monthly time periods. The difference is quite pronounced for 3-motifs of the first set (see, for example, the average percentage of 3-motif number 6 for the monthly networks). Our analysis of the time dependence of the average percentage summarized in figures 5 and 6 indicates that the results obtained for the Bonferroni networks are more robust and reliable than the results obtained for the original network, and this allows for a more detailed investigation of the process of formation and disappearance of these communication structures. In the next section, we will

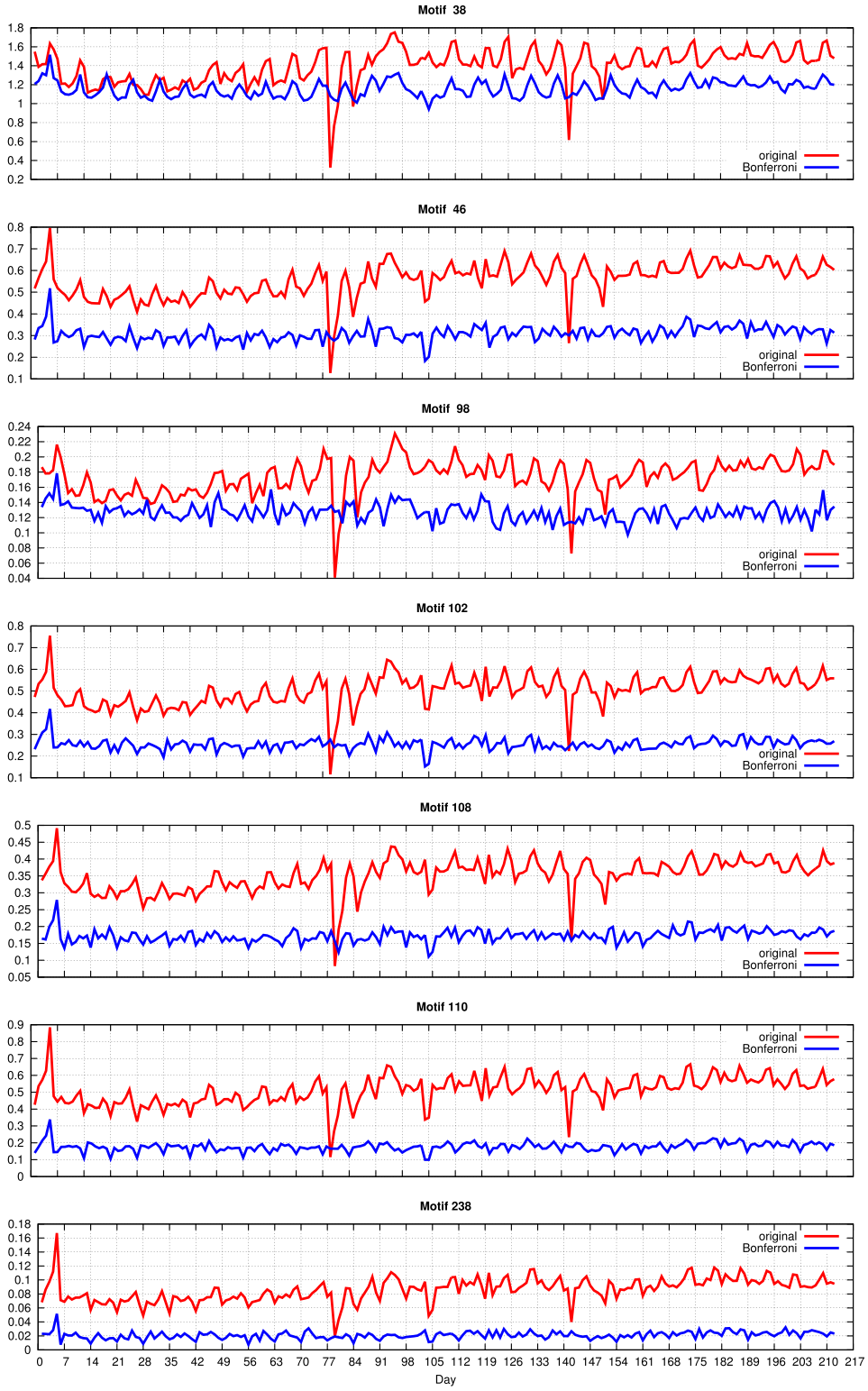

Figure 6. Fraction of 3-motifs observed in each day of the European data set. From top to bottom we have 3-motifs encoded as 38, 46, 98, 102, 108, 110 and 238 respectively. All these 3-motifs have a pair interaction for all pairs in one or both directions during the day investigated. The red line refers to the original network whereas the blue line refers to the Bonferroni network.

analyze the process of formation of the communication 3-motifs in the most reliable setting, which is the setting of the Bonferroni networks.

4. Temporal evolution of communication 3-motifs

Communication 3-motifs are continuously forming and disappearing over time. Here we primarily focus on the dynamics of the 3-motif formation observed at a daily time scale. Specifically, we detect the Bonferroni network at day k and at the two-day time interval beginning at day k, and we count and identify all the 3-motifs present in each network. The identification of each 3-motif is carried out by considering the identity of the three social actors composing it. In other words, we keep a memory of the fact that, for example, one 3-motif of type 6 is observed among subscribers with identities i, j and k. This is done to follow each 3-motif evolution during the increase of the monitoring time interval, which is primarily producing a network expansion[12].

We show in tables 4 and 5 the conditional probabilities for 3-motifs during a one-day expansion of the time period used to determine the Bonferroni network. The starting network is computed for day k, whereas the target network is computed for a two-day time interval including the previous one (days k and $k+1$) for Chinese (table 4) and European (table 5) datasets. The networks refer to the cases of Bonferroni networks obtained from the records of subscribers. We highlight in boldface the entries with conditional probability higher than 0.05. On inspecting tables 4 and 5, we note that the conditional probability $P(M_{\mathrm{II}}|M_{\mathrm{I}})$ shows the highest value in each row when $M_{\mathrm{II}} = M_{\mathrm{I}}$, i.e. when the 3-motif in the expanded network M_{II} is the same as the 3-motif in the starting one M_{I}. This observation suggests that the detection of the 3-motif in the Bonferroni network is pretty robust for all kinds of 3-motifs. The conditional probability ranges between 0.395 (3-motif 98) and 0.916 (3-motif 238) and between 0.418 (3-motif 98) and 0.966 (3-motif 238) for the Chinese and European data respectively. It is worth noting that the less stable 3-motif is the one labeled as 98, which is a motif characterized by a circular flux of information among the three social actors. The second-lowest value of the conditional probability $P(38|38)$ is observed for the other 3-motif which is a triad of unidirectional links.

We also observe that the second-largest value in each row of conditional probabilities is associated with a 3-motif pair requiring that a unidirectional link of the original 3-motif modifies into a bidirectional one in the arrival 3-motif. This observation suggests that the underlying communication process governing the 3-motif dynamics is primarily related to the probability of observing return calls (see $P(14|6)$, $P(14|12)$ and $P(74|12)$, $P(78|14)$, etc) between two social actors.

We provide in figure 7(a) schematic representation of the most relevant conditional probability among the different 3-motifs. We draw a line in the panels of the figure when the conditional probability from the originating to the arrival 3-motif exceeds 5%. For both the Bonferroni networks obtained from the one-day expansion of the Chinese and European datasets, we observe that the typical path of a 3-motif communication does not preferentially

[12] Indeed during the increase of the time interval used to obtain the Bonferroni network of a longer time period, some links existing in the first Bonferroni network might also disappear due to the absence of validation of the link in the second extended period of detection but not in the first period. The probability of disappearance of a link is fairly small, but in a few cases such events occur.

Table 2. Statistics of 3-motifs for the Chinese data. Subscribers only. The networks investigated are the original network and the Bonferroni network. We show the average value observed for real data μ and the average value μ_{rnd} observed by randomly shuffling the network. We also report the standard deviation observed for real data σ and shuffled data σ_{rnd}. Values are given as percentages. Daily (d), weekly (w) and monthly (m) time periods are shown. Values labeled in boldface indicate positive (+) or negative (−) z-score values larger than 3 in absolute value. The z-score is computed as $z = (\mu - \mu_{\text{rnd}})/\sigma$.

		Original				Bonferroni			
		μ	σ	μ_{rnd}	σ_{rnd}	μ	σ	μ_{rnd}	σ_{rnd}
	6	17.41	1.12	17.45	1.03	19.15	0.48	19.3	0.47
	12	23.69	0.85	24.02	0.84	31.32	0.85	31.78	0.82
	14	17.93	0.49	19.16	0.53	14.69	0.36	15.37	0.39
	36	14.09	0.5	14.47	0.47	16.08	0.44	16.43	0.41
	38	**1.09** (+)	0.04	4.22e-5	2.18e-5	**0.94** (+)	0.05	8.81e-6	2.29e-5
	46	**0.57** (+)	0.04	1.66e-5	1.24e-5	**0.31** (+)	0.03	4.31e-6	1.71e-5
d	74	16.08	0.39	17.32	0.45	12.58	0.34	13.28	0.38
	78	6.37	0.44	7.57	0.55	3.45	0.22	3.84	0.25
	98	**0.2** (+)	6.91e-3	7.43e-6	7.64e-6	**0.25** (+)	0.02	1.11e-6	8.15e-6
	102	**0.75** (+)	0.04	2.24e-5	1.52e-5	**0.5** (+)	0.04	5.57e-6	1.96e-5
	108	**0.49** (+)	0.04	1.36e-5	1.12e-5	**0.26** (+)	0.03	5.34e-7	5.58e-6
	110	**1.05** (+)	0.1	2.43e-5	1.58e-5	**0.4** (+)	0.04	3.88e-6	1.7e-5
	238	**0.28** (+)	0.03	1.37e-6	3.54e-6	**0.07** (+)	0.01	1.03e-6	1.07e-5
	6	14.42	3.06	14.13	2.76	12.07	0.46	12.16	0.46
	12	16.89	0.97	16.91	0.95	20.36	0.85	20.33	0.87
	14	20.18	0.73	21.62	0.69	**20.33** (−)	0.3	21.66	0.33
	36	11.08	0.65	11.16	0.63	12.44	0.31	12.56	0.33
	38	**0.69** (+)	0.04	7.91e-5	2.34e-5	**0.78** (+)	0.02	2.63e-5	3.79e-6
	46	**0.59** (+)	0.02	3.5e-5	5.01e-6	**0.58** (+)	0.02	1.34e-5	3.25e-6
w	74	20.02	0.78	21.39	0.73	**19.6** (−)	0.41	20.85	0.44
	78	**12.01** (−)	0.65	14.8	0.78	**10.29** (−)	0.68	12.43	0.87
	98	**0.13** (+)	8.26e-3	1.17e-5	1.7e-6	**0.15** (+)	7.93e-3	5.72e-6	1.5e-6
	102	**0.77** (+)	0.02	5.32e-5	6.09e-6	**0.75** (+)	0.02	2.09e-5	3.27e-6
	108	**0.53** (+)	0.02	4.39e-5	1.31e-5	**0.5** (+)	0.02	1.16e-5	3.31e-6
	110	**1.83** (+)	0.09	8.51e-5	1.11e-5	**1.55** (+)	0.12	2.89e-5	6.32e-6
	238	**0.85** (+)	0.07	1.19e-6	6.02e-7	**0.61** (+)	0.07	3.06e-7	3.36e-7
	6	17.29	5.86	16.47	5.36	8.68	0.23	8.67	0.22
	12	13.08	0.79	12.95	0.75	14.54	0.28	14.24	0.29
	14	19.46	1.41	20.74	1.36	**21.65** (−)	0.1	22.92	0.07
	36	9.44	0.8	9.35	0.75	10.45	0.12	10.31	0.11
	38	**0.4** (+)	0.04	1.65e-4	6.5e-5	**0.49** (+)	0.01	3.39e-5	3.9e-6
	46	**0.46** (+)	0.03	5.79e-5	5.72e-6	**0.57** (+)	0.01	2.59e-5	1.74e-6
m	74	20.56	1.68	21.69	1.6	**22.47** (−)	0.16	23.55	0.15
	78	**15.12** (−)	1.02	18.79	1.03	**16.35** (−)	0.34	20.3	0.45
	98	**0.07** (+)	6.88e-3	1.49e-5	1.94e-6	**0.07** (+)	2.48e-3	6.23e-6	8.99e-7
	102	**0.59** (+)	0.04	8.38e-5	8.21e-6	**0.65** (+)	0.02	3.61e-5	4.85e-6
	108	**0.43** (+)	0.03	1.11e-4	4.07e-5	**0.51** (+)	0.01	2.24e-5	2.92e-6
	110	**1.88** (+)	0.09	1.71e-4	1.33e-5	**2.16** (+)	0.07	7.41e-5	6.86e-6
	238	**1.23** (+)	0.04	3.16e-6	1.67e-6	**1.39** (+)	0.06	1.12e-6	6.83e-7

Table 3. Statistics of 3-motif European data. Subscribers only. The networks investigated are the original network and the Bonferroni network. We show the average value observed for real data μ and the average value μ_{rnd} observed by randomly shuffling the network. We also report the standard deviation observed for real data σ and shuffled data σ_{rnd}. Values are given as percentages. Daily (d), weekly (w) and monthly (m) time periods are shown. Values labeled in boldface indicate positive (+) or negative (−) z-score values larger than 3 in absolute value. The z-score is computed as $z = (\mu - \mu_{\text{rnd}})/\sigma$.

		Original				Bonferroni			
		μ	σ	μ_{rnd}	σ_{rnd}	μ	σ	μ_{rnd}	σ_{rnd}
d	6	25.2	5.46	25.16	5.16	26.67	0.98	26.81	0.97
	12	22.59	2.36	23.33	2.32	28.49	0.45	29	0.46
	14	15.94	1.93	16.84	1.96	14.16	0.74	14.56	0.77
	36	17.31	6.4	18.01	6.14	17.22	0.33	17.89	0.32
	38	**1.4 (+)**	0.19	5.87e-6	2.89e-6	**1.16 (+)**	0.08	1.68e-6	2.76e-6
	46	**0.56 (+)**	0.08	1.45e-6	9.91e-7	**0.31 (+)**	0.03	3.41e-7	1.28e-6
	74	11.53	1.34	12.35	1.38	9.06	0.53	9.42	0.55
	78	3.85	0.57	4.3	0.62	2.19	0.21	2.33	0.23
	98	**0.17 (+)**	0.02	6.68e-7	5.77e-7	**0.13 (+)**	0.01	1.07e-7	7.06e-7
	102	**0.5 (+)**	0.07	1.21e-6	8.6e-7	**0.25 (+)**	0.03	3.56e-7	1.38e-6
	108	**0.35 (+)**	0.05	1.16e-6	8.54e-7	**0.17 (+)**	0.02	3.94e-7	1.39e-6
	110	**0.51 (+)**	0.09	9.31e-7	7.4e-7	**0.18 (+)**	0.03	1.88e-7	1.04e-6
	238	**0.09 (+)**	0.02	9.92e-8	2.22e-7	**0.02 (+)**	5.27e-3	2.06e-8	3e-7
w	6	21.9	8.94	21.25	8.28	16.39	0.28	16.48	0.25
	12	14.33	2.07	14.6	1.93	17.61	0.25	17.97	0.27
	14	19.07	3.34	20.82	3.35	**22.79 (−)**	0.25	24.34	0.27
	36	16.67	7.28	16.82	6.78	**12.64 (−)**	0.29	13.57	0.26
	38	**0.93 (+)**	0.15	1.75e-5	1.05e-5	**1.4 (+)**	0.05	4.42e-6	9.2e-7
	46	**0.88 (+)**	0.16	5.76e-6	2.14e-6	**1.1 (+)**	0.05	2.15e-6	3.95e-7
	74	14.45	2.12	15.94	2.17	**15.84 (−)**	0.27	17.19	0.26
	78	8.47	1.55	10.57	1.81	**8.97 (−)**	0.22	10.45	0.31
	98	**0.08 (+)**	0.01	8.51e-7	2.92e-7	**0.11 (+)**	7.86e-3	5.13e-7	2.87e-7
	102	**0.62 (+)**	0.1	3.83e-6	8.96e-7	**0.71 (+)**	0.04	2.11e-6	7.73e-7
	108	**0.5 (+)**	0.08	4.23e-6	1.6e-6	**0.59 (+)**	0.03	1.24e-6	5.95e-7
	110	**1.53 (+)**	0.28	5.21e-6	1.21e-6	**1.44 (+)**	0.09	2.51e-6	8.24e-7
	238	**0.56 (+)**	0.12	1.05e-6	2.91e-7	**0.41 (+)**	0.04	4.91e-7	2.71e-7
m	6	21.02	12.34	20.29	11.68	12	0.33	11.8	0.3
	12	10.84	2.16	10.82	1.98	13.18	0.17	12.97	0.2
	14	17.47	4.57	18.91	4.57	**24.26 (−)**	0.33	25.99	0.3
	36	22.28	8.34	21.58	7.78	11.06	0.26	11.46	0.25
	38	**0.45 (+)**	0.11	5.54e-5	3.54e-5	**0.84 (+)**	0.02	6.84e-6	2.94e-7
	46	**0.63 (+)**	0.16	1.63e-5	6.5e-6	**1.16 (+)**	0.03	4.97e-6	5.34e-7
	74	14.29	2.86	15.5	2.9	**18.11 (−)**	0.41	19.63	0.45
	78	9.93	2.63	12.89	3.15	**14.37 (−)**	0.27	18.15	0.41
	98	**0.03 (+)**	6.22e-3	1.26e-6	3.08e-7	**0.05 (+)**	4.71e-3	7.55e-7	1.45e-7
	102	**0.38 (+)**	0.09	6.82e-6	1.9e-6	**0.62 (+)**	0.04	4.34e-6	4.34e-7
	108	**0.35 (+)**	0.08	1.03e-5	5.55e-6	**0.61 (+)**	0.03	2.85e-6	2.6e-7
	110	**1.47 (+)**	0.36	1.2e-5	3.23e-6	**2.43 (+)**	0.1	7.79e-6	7.9e-7
	238	**0.87 (+)**	0.23	3.28e-6	9.42e-7	**1.3 (+)**	0.07	2.3e-6	3.48e-7

Table 4. Conditional probabilities for 3-motifs during one-day expansion of the time window used to determine a daily network. The starting network is computed for day k, whereas the target network is computed for a two-day time interval including the previous one (days k and $k + 1$). Motifs detected in the target network are given in columns, whereas motifs detected in the starting networks are given in rows. Chinese data. Subscribers only, Bonferroni networks.

	6	12	14	36	38	46	74	78	98	102	108	110	238	
														None
6	**0.626**	0.004	**0.228**	0.001	0.02	0.006	0.002	0.026	0	0.005	0.005	0.004	0.001	**0.073**
12	0.003	**0.593**	**0.130**	0.002	0.011	0.004	**0.121**	0.032	0.006	0.008	0.004	0.005	0	**0.081**
14	0.012	0.013	**0.711**	0	0.001	0.018	0.007	**0.163**	0	0.014	0.001	0.02	0.004	0.037
36	0.001	0.006	0.002	**0.624**	0.019	0.005	**0.225**	0.028	0	0.005	0.005	0.005	0.001	**0.076**
38	0.016	0.006	0.016	0.017	**0.444**	**0.125**	0.012	0.007	0.001	**0.104**	**0.130**	**0.098**	0.019	0.005
46	0.004	0.002	0.032	0	0.008	**0.538**	0.004	0.012	0	0.004	0.004	**0.316**	**0.076**	0
74	0	0.013	0.007	0.012	0	0.001	**0.702**	**0.171**	0	0.014	0.014	0.021	0.004	0.04
78	0.001	0.001	0.029	0	0	0.001	0.027	**0.858**	0	0.003	0.001	**0.057**	0.018	0.005
98	0	0.041	0.025	0	0.008	0.003	0.022	0.005	**0.395**	**0.321**	0	**0.142**	0.033	0.005
102	0	0.001	0.013	0	0.006	0.01	0.012	0.022	0.003	**0.463**	0.016	**0.371**	**0.080**	0.003
108	0.002	0	0	0.002	0.01	0.01	0.034	0.007	0	0.012	**0.517**	**0.333**	**0.071**	0
110	0	0	0.001	0	0.003	0.012	0.001	0.029	0	0.02	0.009	**0.622**	**0.303**	0
238	0	0	0	0	0	0	0	0	0	0	0	**0.084**	**0.916**	0

Table 5. Conditional probabilities for 3-motifs during one-day expansion of the time window used to determine a daily network. The starting network is computed for day k, whereas the target network is computed for a two-day time interval including the previous one (days k and $k + 1$). Motifs detected in the target network are given in columns, whereas motifs detected in the starting networks are given in rows. European data. Subscribers only, Bonferroni networks.

	6	12	14	36	38	46	74	78	98	102	108	110	238	None
6	**0.707**	0.001	**0.220**	0	0.021	0.006	0	0.020	0	0.003	0.003	0.002	0	0.016
12	0	**0.636**	**0.170**	0	0.017	0.005	**0.114**	0.03	0.004	0.006	0.003	0.004	0	0.010
14	0.002	0.004	**0.806**	0	0	0.022	0.001	**0.133**	0	0.008	0	0.012	0.001	0.011
36	0	0.001	0	**0.706**	0.032	0.006	**0.200**	0.026	0	0.003	0.005	0.003	0	0.017
38	0.003	0.007	0.004	0.005	**0.601**	**0.139**	0.003	0.002	0.001	**0.067**	**0.099**	**0.059**	0.009	0.001
46	0	0	0.024	0	0.006	**0.731**	0	0.006	0	0.004	0	**0.194**	0.032	0.002
74	0	0.002	0.002	0.001	0	0	**0.758**	**0.180**	0	0.012	0.017	0.014	0.002	0.011
78	0	0	0.008	0	0	0	0.006	**0.922**	0	0.001	0	0.048	0.014	0.001
98	0	0.005	0.005	0	0	0	0	0.005	**0.418**	**0.409**	0	**0.132**	0.027	0
102	0	0	0.015	0	0.002	0	0.010	0.005	0	**0.521**	0	**0.388**	**0.059**	0
108	0	0.004	0	0	0.007	0	0.014	0.014	0	0.004	**0.634**	**0.276**	0.043	0.004
110	0	0	0.004	0	0	0	0	0.021	0	0	0.008	**0.736**	**0.230**	0
238	0	0	0	0	0	0	0	0	0	0	0	0.034	**0.966**	0

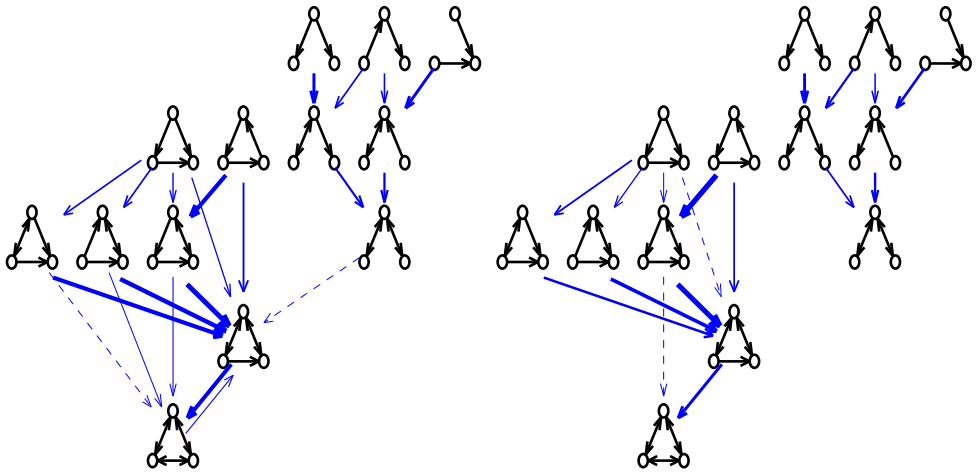

Figure 7. Schematic representation of transition probabilities of 3-motifs when the network expands from a Monday network to a Monday–Tuesday network. Transition probabilities of more than 0.05 are indicated with blue arrows. The thickness of an arrow is proportional to the value of the conditional probability. The left and right panels refer to the Chinese and European datasets respectively.

show triadic closure of open triangles, but rather completion of reciprocal calls. In other words, the typical evolution path of a 3-motif, when the time interval used to detect the network extends from one day to two days, shows that just a small fraction of 3-motifs evolves from open 3-motifs (i.e. 3-motifs with communication links detected only between two pairs of actors) to closed 3-motifs (i.e. 3-motifs with communication links detected between all three pairs of actors). Open 3-motifs preferentially tend to evolve towards bidirectional open 3-motifs, and only when links are fully reciprocated in the open 3-motif (motif 78) does the motif tend to evolve to a closed 3-motif—along the lines of triadic closure.

We interpret this observation as a manifestation of the fact that communication closed 3-motifs typically form at an intraday time scale. This interpretation is also supported by the results reported in figure 8, where we look at the details of formation and evolution of closed 3-motifs from one day to the next one, without varying the time window and without distinguishing among the different closed 3-motifs and the different open 3-motifs. On average, more than 2/3 of the closed 3-motifs observed in the Bonferroni network at a given day come from unconnected triples of nodes in the Bonferroni network of the previous day, and evolve to unconnected triplets of nodes in the Bonferroni network of the following day (red rectangles in the figure). On the other hand, the closed 3-motifs originating from (evolving to) open 3-motifs in the Bonferroni network of the previous (following) day amount to about 1/4 of the total (blue triangles in the figure). Such an erratic pattern suggests that a large fraction of closed 3-motifs that appear in a daily Bonferroni network occur due to contingent reasons of communication that develop at an intraday time scale, e.g., the peak of closed 3-motifs observed on Friday may be due to the need for people to coordinate their social activities.

5. Conclusions

In this paper, we have adapted and applied a filtering procedure to a directed communication network. This filtering procedure is based on a statistical validation performed by using

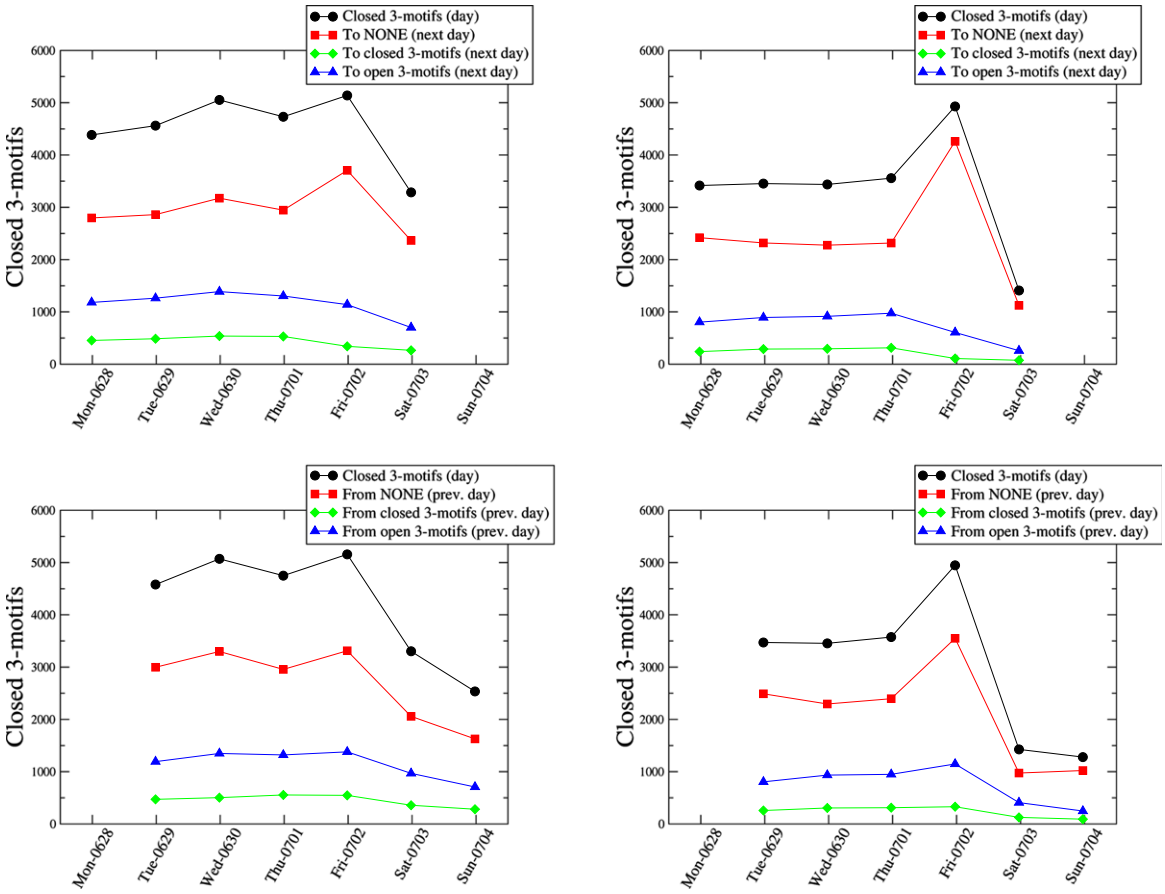

Figure 8. Top panels: evolution of closed 3-motifs for Chinese (left) and European (right) data across a week. Black circles indicate the total count of closed 3-motifs in the daily Bonferroni networks for the first six days of a week. Red rectangles indicate the number of these motifs evolving, the day after, to node triplets that do not determine a 3-motif. Blue triangles (green diamonds) indicate the number of closed 3-motifs evolving, the day after, to open 3-motifs (closed 3-motifs). Bottom panels: formation of closed 3-motifs for Chinese (left) and European (right) data across a week. Black circles indicate the total count of closed 3-motifs in the daily Bonferroni networks for the last six days of a week. Red rectangles indicate the number of these motifs emerging from node triplets that did not determine a 3-motif in the Bonferroni network of the day before. Blue triangles (green diamonds) indicate the number of closed 3-motifs that were open 3-motifs (closed 3-motifs) in the Bonferroni network of the day before.

multiple-hypothesis test correction. We hypothesize that the links detected in the directed Bonferroni communication networks describe relevant ties of the underlying social structure originating the communication. We test our hypothesis by comparing basic statistics of the original network and the Bonferroni network, and conclude that the latter is much more realistic as it removes spurious links related to non-social interactions. Furthermore, we investigate the relative frequency of 3-motifs in two large sets of mobile communication data recorded in two different countries of two distinct continents. In both cases, we verify that the frequency profile of the 3-motifs of the directed Bonferroni communication networks is much more stable over time than the frequency profile of the original network. We believe that this empirical

observation supports the hypothesis of Bonferroni networks being good proxies of strong ties of social origin.

After having verified the robustness and reliability of our statistical filtering procedure, we investigated the time evolution of the communication 3-motifs. Our results show that communication 3-motifs characterized by triadic closure form frequently at an intraday time scale. On the other hand, open 3-motifs (i.e. 3-motifs with links detected only between two pairs of subscribers) primarily evolve to other open 3-motifs with a higher number of reciprocated calls. In fact, the preferential path of evolution of open 3-motifs shows that an open 3-motif evolves to a closed triad with a sizable conditional probability after all the calls of the open 3-motif are reciprocated.

We interpret these results as evidence for the fact that correctly sampled mobile call records reflect rapid communication interactions of an underlying social structure that forms and dissolves over a longer time scale. In other words, the time scales of the communication network and of the social network are quite distinct, with the first lasting usually less than one day and the second requiring months or years. Under this interpretation, we conclude that the triadic closure process is governed by distinct rules in communication and in social networks.

Acknowledgments

The authors thank L Barabási for the European data. This work was partially supported by the National Natural Science Foundation of China (11205057), the Humanities and Social Sciences Fund of the Ministry of Education of China (09YJCZH040), the PhD Programs Foundation of the Ministry of Education of China (20120074120028), the Fok Ying Tong Education Foundation (132013), and the Fundamental Research Funds for the Central Universities.

References

[1] Lazer D *et al* 2009 Computational social science *Science* **323** 721–3
[2] Onnela J-P *et al* 2007 Structure and tie strengths in mobile communication networks *Proc. Natl Acad. Sci.* **104** 7332–6
[3] Eagle N, Pentland A and Lazer D 2009 Inferring friendship network structure by using mobile phone data *Proc. Natl Acad. Sci.* **106** 15274–8
[4] Palla G, Barabási A-L and Vicsek T 2007 Quantifying social group evolution *Nature* **446** 664–7
[5] Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E 2008 Fast unfolding of community hierarchies in large networks *J. Stat. Mech.* P10008
[6] Song C, Qu Z, Blumm N and Barabási A-L 2010 Limits of predictability in human mobility *Science* **327** 1018–21
[7] Lu X, Bengtsson L and Holme P 2012 Predictability of population displacement after the 2010 Haiti earthquake *Proc. Natl Acad. Sci.* **109** 11576–81
[8] Schneider C M, Belik V, Couronné T, Smoreda Z and González M C 2013 Unravelling daily human mobility motifs *J. R. Soc. Interface* **10** 20130246
[9] Karsai M, Kivelä M, Pan R K, Kaski K, Kertész J, Barabási A-L and Saramäki J 2011 Small but slow world: How network topology and burstiness slow down spreading *Phys. Rev.* E **83** 025102
[10] Kivelä M, Pan R K, Kaski K, Kertész J, Saramäki J and Kasrai M 2012 Multiscale analysis of spreading in a large communication network *J. Stat. Mech.* P03005

[11] Kovanen L, Karsai M, Kaski K, Kertész J and Saramäki J 2011 Temporal motifs in time-dependent networks *J. Stat. Mech.* P11005
[12] Jiang Z-Q, Xie W J, Li M-X, Podobnik B, Zhou W-X and Stanley H E 2013 Calling patterns in human communication dynamics *Proc. Natl Acad. Sci.* **110** 1600–5
[13] Blondel V D *et al* 2012 Data for development: the D4D challenge on mobile phone data arXiv:1210.0137
[14] Granovetter M 1973 The strength of weak ties *Am. J. Sociol.* **78** 1360–80
[15] Barabási A-L 2005 The origin of bursts and heavy tails in human dynamics *Nature* **435** 207–11
[16] Onnela J-P *et al* 2007 Analysis of a large-scale weighted network of one-to-one human communication *New J. Phys.* **9** 179
[17] Palchykov V, Kaski K, Kertész J, Barabási A-L and Dunbar R I M 2012 Sex differences in intimate relationships *Sci. Rep.* **2** 370
[18] Kovanen L, Kaski K, Kertész J and Saramäki J 2013 Temporal motifs reveal homophily, gender-specific patterns and group talk in mobile communication networks *Proc. Natl Acad. Sci.* **110** 18070–5
[19] 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
[20] 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
[21] Schneider C M, Belik V, Couronné T, Smoreda Z and González M C 2013 Unravelling daily human mobility motifs *J. R. Soc. Interface* **10** 1742
[22] Wasserman S and Faust K 1994 *Social Network Analysis: Methods and Applications* (Cambridge: Cambridge University Press)
[23] Squartini T, van Lelyveld I and Garlaschelli D 2013 Early-warning signals of topological collapse in interbank networks *Sci. Rep.* **3** 3357
[24] Bargigli L, di Iasio G, Infante L, Lillo F and Pierobon F 2013 The multiplex structure of interbank networks *Quantitative Finance* at press, available at http://ssrn.com/abstract=2352787
[25] Holme P and Saramäki J 2013 Temporal networks *Phys. Rep.* **519** 97–125
[26] Radicchi F, Ramasco J J and Fortunato S 2011 Information filtering in complex weighted networks *Phys. Rev.* E **83** 046101
[27] Serrano M A, Boguñá M and Vespignani A 2009 Extracting the multiscale backbone of complex weighted networks *Proc. Natl Acad. Sci.* **106** 6483–8
[28] Mantegna R N 1999 Hierarchical structure in financial markets *Eur. Phys. J.* **11** 193–7
[29] Baba K, Shibata R and Sibuya M 2004 Partial correlation and conditional correlation as measures of conditional independence *Aust. NZ J. Stat.* **46** 657–64
[30] Aste T, di Matteo T and Hyde S T 2005 Complex networks on hyperbolic surfaces *Physica* A **346** 20–26
[31] Tumminello M, Aste T, di Matteo T and Mantegna R N 2005 A tool for filtering information in complex systems *Proc Natl. Acad. Sci. USA* **102** 10421–6
[32] Hatzopoulos V, Iori G, Mantegna R N, Micciché S and Tumminello M 2013 Quantifying preferential trading in the e-MID interbank market, available at http://ssrn.com/abstract=2343647
[33] Tumminello T, Micciché S, Lillo F, Piilo J and Mantegna R N 2011 Statistically validated networks in bipartite complex systems *PLoS ONE* **6** e17994
[34] Tumminello M, Lillo F, Piilo J and Mantegna R N 2012 Identification of clusters of investors from their real trading activity in a financial market *New J. Phys.* **14** 013041
[35] Tumminello M, Edling C, Liljeros F, Mantegna R N and Sarnecki J 2013 The phenomenology of specialization of criminal suspects *PLoS ONE* **8** e64703
[36] Dunbar R I M 1998 The social brain hypothesis *Evol. Anthropol.* **6** 178–90
[37] Wernicke S 2005 A faster algorithm for detecting network motifs *Proc. 5th Workshop on Algorithms in Bioinformatics (WABI '05)* (*Lecture Notes in Bioinformatics* vol 3692) (Berlin: Springer) pp 165–77