Urban communications and social interactions through the lens of mobile phone data

Sabrina Gaito, Christian Quadri, Gian Paolo Rossi, Matteo Zignani*

Computer Science Department, University of Milan, Milan, Italy

ARTICLE INFO

ABSTRACT

The social network built on top of mobile phone data has drawn increasing attention in recent years, due to its being far more accurate than its online counterpart in mirroring people's offline sociality. In this paper, we leverage a large dataset of mobile Call Detail Records (CDRs), provided by one of the primary Italian operators, to build the multiplex social network given by call and text message activities of the operator's subscribers. By discussing the multiplex network characteristics – ranging from ego-networks, dyads and triads all the way to cliques – the paper offers a comprehensive and thorough overview of human sociality as carried on through mobile phone; in addition, it highlights the need to consider more than one communication media when aiming to understand people's sociality.Finally, by investigating on-phone cliques, we show people's inclination to gather in cohesive and restricted groups of close friends, thus providing strong ominous indicators of where many recent online social networks, namely Snapchat, WeChat and others, are leading.

1. Introduction

Comprehension of the human interaction-social networks has been constrained in accuracy, breadth, and depth because of its dependency on incomplete, biased and uncertain data. For a long time, social sciences have studied human sociality on the basis of self-reported survey data. Their invaluable characteristic of describing people's offline social relations, now almost completely unobservable by means of automatic recording systems, has allowed sociologists to build the core of current sociological theories [1].

Continual development of digital communications, epitomized in the shift from the informational web (Web 1.0) to the interactional web (Web 2.0), is presently generating new opportunities and challenges in the sphere of social research [2]. Nonetheless, both proliferation of Web 2.0 social media and increased web accessibility through smartphones generate new data of great value for social research [3,4]. On the one hand, in fact, their characteristic of being big and broad allows researchers to investigate a wider population sample [5–7], while, on the other, they refer to a kind of sociality that has gradually shifted away from the off-line sociality they sought initially to mirror [8].

The complementarity between online and offline sociality [9,10], as well as between big digital data and survey data, has

become the subject of an intense and still active debate. It has recently turned out that a promising synthesis between the two approaches can be achieved by studying social networks built on top of on-phone interactions [11–14]. Mobile phones are undoubtedly the most common form of device-mediated communication. We have adopted them as a safety link, to communicate within the family, in the personal circle of friendships and at work. This way, on-phone social networks are much more accurate than their online counterpart for mirroring people's offline sociality, all while sharing with the latter the power to provide big and broad data [15–17].

In this paper we leverage a large dataset of mobile Call Detail Records (CDRs) [18], provided by one of the primary Italian operators, to build the multiplex [19–22] social network given by call and text message activities of the operator's subscribers [23]. The paper represents a further step forward in a three-year research on the subject, with the aim of offering a comprehensive and thorough overview of human sociality as carried on by mobile phones [24,25]. It describes and discusses the multiplex network's characteristics, ranging from ego-networks, dyads and triads all the way to cliques. Also, it provides further insights into the dynamics of people's sociality within the context of Dunbar's theory [8,26].

Finally, by showing the social network built on top of people's real-life activities, this paper provides strong ominous indicators of where many recent online social networks are leading. In fact, after an initial period where the priority was to acquire as many

* Corresponding author.
E-mail address: matteo.zignani@unimi.it (M. Zignani).

'friends' as possible, we are now witnessing the rise of a new type of online social service that is more oriented toward playful amusement and intimacy, and less to information and self promotion. Some emerging social networks such as Snapchat, WeChat and Line, prove the point: users like to send goofy images to a restricted group of close friends, a social circle in which they feel comfortable making a fool of themselves, with no need to reflect carefully before publishing a post [27]. Echoing Dunbar's grooming coalition, the trend is taking a direction that can be found in the evidence we provide about cliques and the nature of their inner interactions.

The main contributions provided by our work rely on the multiplex approach we adopt to analyze the social structure of the interactions mediated by on-phone communications. In fact, we argue that a multidimensional view on mobile phone datasets is mandatory to really understand the communication attitudes of phone users. The novel insights provided by the multiplex viewpoint involve all the levels of our network analysis – from the fundamental elements like node and link sets, to the mesoscopical perspective provided by strongly cohesive groups. In comparison to the standard analysis based only on call interactions, the multiplex approach enables us to enlarge the set of users to include in our analysis. To be sure, we found that some users prefer just one form of communication to interact with other people, while the remaining ones combine use of various communication channels (Section 4.1). This phenomenon also characterizes usage of communication channels in maintaining the relationships (Section 4.2). Specifically, half of the interacting pairs adopt exclusively a single channel to communicate. Application of the multidimensional approach in studying of the neighborhood of nodes allows us to verify whether a user maintains her/his level of activity across the diverse dimensions or, conversely, if a user who is active through voice calls is not disposed to communicate by texting. Through the study of the correlation among the different communication layers (Section 6) we found that (*i*) people who contact many other people are expected to be contacted by many individuals; (*ii*) users important in one layer in terms of number of people they contact, and vice versa, may lose their importance in the complementary one (i.e., activity and popularity are not directly maintained across on-phone communication layers); and (*iii*) the number of very active people who interact with most of their neighbors through a single channel is negligible, while they prefer multi-channel communications to interact with most of their neighbors. Through the multiplex viewpoint we also identified the support cliques of the users (Section 7), i.e. the set of the strongest ties; in addition, we found that voice calls represent the main medium used to interact with individuals people are closest to. Voice calls also play an important role in the formation and support of strongly cohesive groups, as in cliques (Section 8). Specifically, interactions within cliques are mainly facilitated by call and multichannel communications.

2. Dataset

This paper is based on a large anonymized dataset of Call Detail Records (CDRs) [28] concerning voice calls, short text messages (SMS) and Internet traffic of about 1 million subscribers of an Italian mobile operator in the metropolitan area of Milan for a period of 67 days (namely, March 26 to May 31, 2012). During this period a total of over 63 million phone calls and 20 million text messages were recorded. The time window covered by the dataset is long enough to reconstruct most of on-phone social relationships, as observed in Onnela et al. [11], where the authors show that the statistical characteristics of the network largely saturate in a two-month sample. The large amount of communication records allows us (*i*)

to study how people maintain their social relationships in an urban context and (*ii*) to investigate the relation between the sociality captured by the combined usage of different communication media (phone call and text messages) and the sociality expressed by a single communication channel.

2.1. Data collection methodology

Cellular network operators implement a set of logical charging services to keep track of the network resource usage by their customers for billing and inter-accounting purposes. Whenever a user makes a call or send a text message, the probing system retrieves data about the sender and the receiver, the cell ID of the handling towers, and the start time-stamp of the activity. Then it stores them in a CDR. Specifically, each entry of a CDR is formally described by the 5-ple $t_{CDR} = \langle s, r, t_{start}, d, loc \rangle$, where s and r, respectively, represent the sender ID and the receiver ID of the call/SMS,[1] t_{start} is the initial time of the activity (when a call starts or an SMS is sent), d is the duration and loc is the serving cell the user s is attached to when the activity gets started. In Fig. 1, we report a small sample for each kind of activity. To bring each record back to the 5-ple, we discard the field 'dir' since it indicates the direction of flow w.r.t. the operator network, so it is useful for the identification of the customers. As for the localization of the user, here we keep only the 'start cell'.

2.2. Extraction of relevant on-phone communications

Text message and phone call records are the main data source for purpose of understanding the social structure of the interactions mediated by on-phone communications, especially before the era of OTT services. Such a complex structure is often represented by a network whose nodes are users and where links connect two users who call/text each other. However, the choice of drawing a link depends on the purpose of the communication. In fact, all calls and texts do not have the same social value; this is particularly true in the case of advertisements and commercial messages or communications issued by call centers. Moreover, in our dataset we observe a curious behavior: nearly 40% of calls have duration equal to 0. Not counting authentically missed or unanswered calls, such a large amount of empty ringing echoes a common practice in Italy to use rings to mean "Call me back soon" or "I'll be there in a sec" – for instance, in synchronizing times for a meeting. Due to the difficulty in discriminating 0-duration calls on the base of their social meaning, in this work we decided to remove these records from the dataset. So, in the end our dataset consisted of 41 million phone calls and 20 million text messages.

Usually, the literature on mobile phone cleansing [4,29–31] suggests filtering out incoming and outgoing communications that involve other mobile operators' customers, so as to eliminate the bias between operators; in fact, we have full access to the call/SMS records of one operator but partial access to calls to/from subscribers of other operators. Here, we apply a hybrid approach: when we deal with users' neighborhood we also exploit the communications from/to other mobile operators' customers (see Sections 5–7). Meanwhile in the analysis of more complex network structures, like cliques (see Section 8), we apply the aforementioned sampling method.

On the basis of the obtained CDR dataset, we construct two preliminary on-phone social networks, one for each communication channel, from which to extract only the interactions with social relevance. To this end, in the call network, we consider the pairs of

---

[1] For the purpose of ensuring customer anonymity, each subscriber is identified by a surrogate key.