Communication Motifs: A Tool to Characterize Social Communications

Qiankun Zhao[♭] Yuan Tian[†] Qi He[†] Nuria Oliver[♭] Ruoming Jin[‡] Wang-Chien Lee[†]
[♭] Telefonica Research, [†] Penn State University, [‡] Kent State University
qkzhao@gmail.com, yxt144@cse.psu.edu, qhe@ist.psu.edu, nuriao@tid.es,
jin@cs.kent.edu, wlee@cse.psu.edu

ABSTRACT

Social networks mediate not only the relations between entities, but also the patterns of information propagation among them and their communication behavior. In this paper, we extensively study the temporal annotations (*e.g.*, time stamps and duration) of historical communications in social networks and propose two novel tools – *communication motifs* and *maximum-flow communication motifs* – for characterizations of the patterns of information propagation in social networks. Using these motifs, we verify the following hypothesis in social communication network: 1) the functional behavioral patterns of information propagation within both social networks are stable over time; 2) the patterns of information propagation in synchronous and asynchronous social networks are different and sensitive to the cost of communication; and 3) the speed and the amount of information that is propagated through a network are correlated and dependent on individual profiles.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval: Online Information Services; Web-based services;**]:

General Terms

Algorithms, Design, Experimentation

Keywords

Social Networks, Information Flows, Motif

1. INTRODUCTION

Social networks represent the links between a set of entities connected to each other with different types of relationships. In the literature, social networks have been extensively studied from a graph theory perspective (*e.g.*, power laws, small worlds phenomenon, coverage, etc.) [1]. Properties of different types of complex networks have been compared [4]. Recently, research studies on social networks have a lot of applications in recommender systems, social search, economics, and advertising [2].

In this paper, we take the behavioral aspect to study the communication characteristics within social networks. In particular, we aim at identifying topological subgraph structures with frequency and temporal constraints, which we refer to as *motifs*, to characterize communications in social networks.

In traditional social networks such as friends or citation networks, the nature of the relationship is embedded in – or may be easily derived from – the records. However, for these social networks derived from communication logs, it is difficult to properly infer the nature of the relationships due to the multiplicity of reasons in making a call (*e.g.*, business, personal, service, etc.) and the implication of the temporal context. In other words, once one paper cited another, the *cited* relationship between both papers always holds true. However, phone calls are made for different reasons and hence the nature of a relationship between two nodes in the network may also depend on the temporal context of calls, i.e., a call made during working hours is probably of different nature than a call made at night. The same applies to other temporal attributes such as duration and frequency of the interaction, or temporal distance between two calls (inter-call time delay). However, many studies on information propagation assume that consecutive interactions transmit the same piece of information within the inferred networks, which we shall argue that this assumption is not valid later.

In [1], Kossinets *et al.* notice the importance of the temporal annotation in instances of communications. They find an information pathway where users are updated with the latest information at the quickest speed based on the temporal distance between the communications. However, their work does not consider the case of different pieces of information are propagated in similar contexts. Moreover, by using part of the temporal information only, they are not able to generalize the local behavior patterns of individuals or small groups to the entire network.

In this paper, we propose to study information propagation and behavior patterns in communication oriented social networks by leveraging the **temporal annotations** of these communications together with the topological structure of the network. The proposed approach is based on the following observations:

- Calls or interactions between the same two users in the network may have different purposes and thus trans-

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

mit different information, resulting different effects in terms of information propagation. The interactions between any two users may range from being intense and frequent, hence generating a lot of follow-up calls and reaching a lot of users, to being isolated events without further impact on the network. As a result, we propose to differentiate the calls by incorporating temporal information and frequency of the calls into the social networks.

- The semantics associated with each interaction (*e.g.*, topics discussed, purpose of a call) are hard to infer [4]. It is possible that two adjacent interactions (*i.e*„ interactions that share at least one common user) have no causality relationship between them. However, the temporal attributes associated with the interactions may shed some light on information propagated. For example, it seems reasonable to assume that the closer in time two adjacent interactions take place, the more likely it is that they are about the same topic.
- The amount of information passed from one node to another in the social network may be quantified in different ways. For example, in a CDR dataset the amount of information can be quantified by the duration of the call, whereas in a Facebook dataset the amount of information can be quantified by the length of the text typed on a user's wall. In both cases, we assume that the longer an interaction is, the larger amount of information is propagated via this interaction.

In this paper, we propose the concept of *communication motifs* and *maximum-flow motifs* to model how information is propagated in social networks. The proposed motif concepts are an extension of the network motif in biological networks [3], which refers to patterns that recur within a network much more often than at random. By defining the notion of communication motifs, we carry out extensive experiments with two types of social networks (synchronous and asynchronous). The result presents a number of properties exhibited in the discovered motifs. Finally, we characterize the information propagation behavior in these social networks using these motifs as a measurement.

2. COMMUNICATION MOTIFS

Communication motifs in social networks are motivated by the assumption that the communication within the social network is event-driven [4], *e.g.*, each call in a CDR dataset is made with the purpose of propagating or getting a certain piece of information. However, individual users may make and receive a lot of phone calls. In this paper, we shall assume that: (a) *Each interaction between two users and its corresponding timestamp is taken as evidence of information propagation between the users*; (b) *two adjacent interactions (i.e., interactions that share at least one common user) are more likely to be about the same piece of information if they are temporally close*; (c) *every piece of information is only valid for a time window W (e.g., 1 hour or 2 days depending on the type of network)*; and (d) *users either propagate this information via immediate interactions within W or they won't propagate it any more.* In this section, we first define the concept of *communication motif* and *maximum-flow motif* based on the above assumptions. Then, we propose two algorithms for discovering such motifs. Hereafter, we use phone call networks created from Call Detail Records (CDRs) as an example of a communication network.

Let a database of CDRs be denoted by G. Each record in G is represented as 4-tuple e_i=$< u_i^o, u_i^d, t_i, \delta_i >$, where u^o represents the caller, u^d the callee, t_i the timestamp when the call was established, and δ_i the duration of the call. To represent the information flow in this type of network, we define a communication graph as:

Definition 1. **(Communication Graph)** *For a given database of CDR represented as G and a time window W denoting the life span of every piece of information, a communication graph is defined as a subset of records from G, denoted as* $C = \{e_1, e_2, \cdots, e_k\}$*, such that* $\forall e_i \in C$, $1 \leq i \leq k$*, there exists at least one user* $e_j \in C$ *and* $i \neq j$ *such that 1)* $|\{u_i^o, u_i^d\} \cap \{u_j^o, u_j^d\}| > 0$*, and 2)* $0 < t_i - t_j - \delta_j < W$ *or* $0 < t_j - t_i - \delta_i < W$.

Note that even though there are k entries in the communication graph, the number of unique users involved can be different depending on the characteristics of the calls that form this communication graph. Therefore, each communication graph is annotated with two numbers C_i^j, where i denotes the number of unique users (nodes in C) and j denotes the number of interactions (edges) that form this communication graph.

The *temporal* constraint given by W is the core filter to eliminate some of the noise in the data. Note also that the individual communication graph is a simplification of reality. In real data, it is common to observe that a person is involved in multiple communication chains, even during a short time period, since (s)he may receive or initiate multiple calls from or to more than one persons within the predefined time window W.

Next, we introduce the concepts of *communication motifs* based on *graph equivalence.*

Definition 2. **(Communication Motif)** *Given a support threshold* θ*, then any equivalence class of the communication graphs* $S = \{C_i^j(1), C_i^j(2), \cdots, C_i^j(n)\}$ *based on the graph-equivalence with no less than* θ *communication graphs is referred to as a* communication motif. *In addition, the cardinality of the equivalence class S is referred to as the* support *of the communication motif.*

Finally, note that in order to deal with potential biases and local noise, we would like the communication graphs of the motifs to be evenly distributed in the social network, particularly given that the motifs are supposed to capture intrinsic properties of the network. Thus, we introduce the concept of *L-support* to set an upper bound on the number of communication graphs per origin that are taken into account in the overall support of the communication motifs. Note that L-support generalizes to the original support definition: as L becomes larger, the L-support converges to the original support.

In sum, we are interested in the functional communication patterns of social interaction networks that not only occur frequently but also are indicative of the process of information propagation in the network. They are meant to capture patterns on the collective behavior of users in the network. However, if we want to apply the motifs to real-world scenarios such as viral marketing, we need to define a measure to quantify the probability of information propagation within each motif.

Definition 3. **(Maximizing the flow of informa-**

tion) *Given two connected calls* $e_i = < u_i^o, u_i^d, t_i, \delta_i >$ *and* $e_j = < u_j^o, u_j^d, t_j, \delta_j >$, e_i *and* e_j *maximize the flow of information iff for all connected calls* G', *where* $e_j \in G'$ *there is no other* $e_{j'}$ *such that* $\delta_j/(\delta_i \times (t_j - t_i)) \leq \delta_{j'}/(\delta_i \times (t_{j'} - t_i))$.

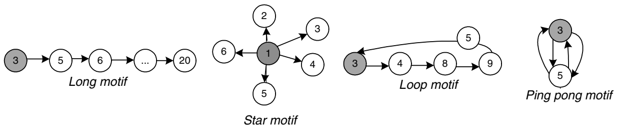


Figure 1: Types of communication motifs.

Based on the previous definitions, the communication motifs may reveal aspects of the principles and dynamics in the communications that take place within the network. Figure 1 illustrates four basic types of communication motifs with different information flow patterns: (a) *Long chain*: which is a communication graph with an extreme long list of *unique* participants, such as 3, 5, 6, $\cdots$,20 in the figure; (b) *Ping pong*: which represents a pattern with very few participants that repeatedly communicate back-and-forth with each other, such as 3, 5 in the figure; (c) *Loop*: where there is a loop in the communication between some of the participants – but not the *ping pong* loops, such as from 3 to 4, 8, 9, and back to 3; and (d) *Star*: where all the interactions start/end on a limited number of members of the graph.

In general, a communication motif involves a number of distinct participants and a combination of the basic motifs defined above. The distribution of the basic motifs within the larger motif is an indication of how information flows within the motif, how fast it may spread and of the topology of the social network. on the motifs that have more than two nodes.

3. EXPERIMENTS

We have conducted a comprehensive analysis of two real social network communication datasets in order to study the characteristics of the motifs within these social networks. In this section, we present the data analysis results and discuss the insights from our findings.

Dataset	*CDR*	*Facebook*	*Random*
Isolated entries	74.3%	69.7%	99.99%
2-people Ping-pong	11.4%	14.5%	0.01%
Communication motifs	14.3%	15.8%	0

Table 1: Overall distribution of motifs ($W = 4$hr)

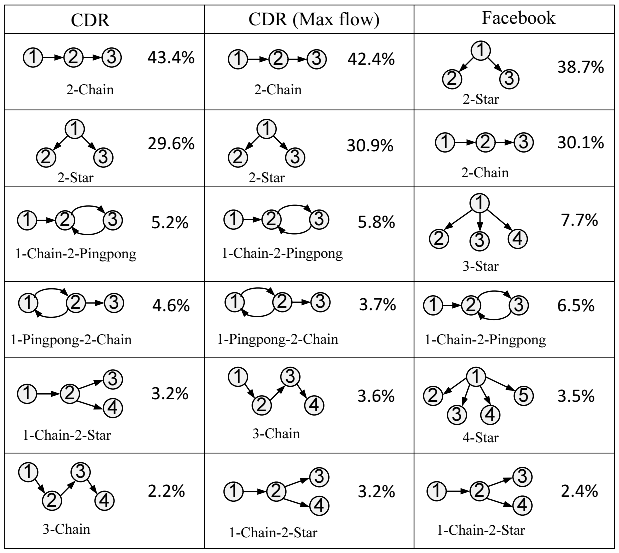
The two real social network communication datasets used in our experiments are:

(1) A *Call Detail Record* (CDR) dataset that has been collected from a large western city for a duration of *3* months and for a specific mobile service provider. It contains *17,800,000* entries (phone calls) among *245,600* unique users.

(2) The wall-post history of Facebook, collected for the New Orleans Facebook network for a duration of over *2* years with *800,000* interactions among *65,000* unique users.

The communication motifs are extracted from both CDR and Facebook, and the maximum information flow motifs are from CDR using the duration of each call.

General properties of motifs



CDR	CDR (Max flow)	Facebook
2-Chain 43.4%	2-Chain 42.4%	2-Star 38.7%
2-Star 29.6%	2-Star 30.9%	2-Chain 30.1%
1-Chain-2-Pingpong 5.2%	1-Chain-2-Pingpong 5.8%	3-Star 7.7%
1-Pingpong-2-Chain 4.6%	1-Pingpong-2-Chain 3.7%	1-Chain-2-Pingpong 6.5%
1-Chain-2-Star 3.2%	3-Chain 3.6%	4-Star 3.5%
3-Chain 2.2%	1-Chain-2-Star 3.2%	1-Chain-2-Star 2.4%

Figure 2: Most frequent motifs ($W = 4$hrs)

Table 1 presents the overall distribution of three types of basic communication building blocks in three networks (CDR, Facebook and random): isolated entries (*i.e.* entries that have no other related entries within the given time window); 2-people Ping-pong motifs (repeated entries that only involve 2 people) and other communication motifs. The *random* network is a synthetic communication network that: (1) has exactly the same topology as the CDR dataset (*i.e.,* the number of calls between any two users is identical), but where (2) the timestamps for all calls that are made within the given time period are randomized. We run the motif extraction algorithms on this synthetic dataset. In all datasets, we can see that the majority of entries are isolated. However, while the the 2-people ping-pong motifs and communication motifs both comprise about 15% of the total entries in the CDR and Facebook datasets, such motifs can hardly be found in random networks. This result strongly suggests that the time-constrained motifs in social communication networks are driven by the special characteristics of human activities (such as the propagation of information).

Motif frequency distribution. First, we extract the motifs from both social networks and compute how frequent they are within each network. Figure 2 shows the 6 most frequent motifs from the datasets (obtained with a time window W = 4 hours). Note that hereafter the support is converted into a percentage, based on the ratio of frequency to the total number of entries in the dataset so that the results across datasets are comparable. Clearly, there are frequent motifs that are shared by both datasets, such as the 2-chain and 2-star topologies. This indicates that the communication behaviors of people are similar even in different environments. On the other hand, we can also observe some differences in the frequent motif lists: the Facebook dataset contains considerably more star structures than the CDR dataset while the CDR dataset has more chains.

Stability of motifs over time

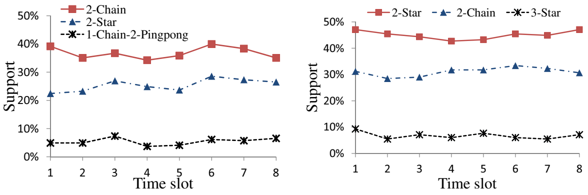
In this section, we evaluate the stability of the discovered motifs from both networks to explore the applicability of using such motifs to characterize communications in social networks, which are often very dynamic over time.

It has already been discussed in Section *4.1* that the motif frequency distribution is not sensitive to the time window size W. To capture the motif distribution changes over time, here we use Kendall's tau coefficient to measure the similarity of two ranked motif lists L and L' in different time slots, as given by Equation 1.

$$\tau(L, L') = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \qquad (1)$$

where n is the length of L and L', n_c is the number of concordant pairs in L and L' and n_d is the number of discordant pairs in the lists. If $\tau(L, L') = 1$, that means the two lists are identical, while $\tau(L, L') = -1$ means there the rankings of the two lists are totally different (*e.g.*, reversed).

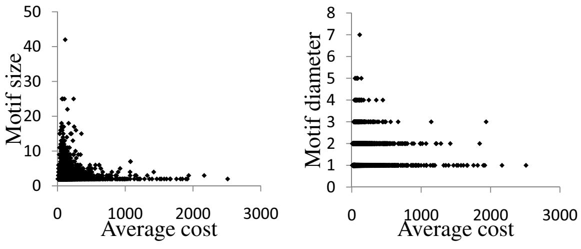
Figure 3 shows the evaluation results. We partition the entire time period of the datasets (3 months for the CDR dataset and 2 years for the Facebook dataset) into eight equal-length time bins (slots) and extract the motifs on the subset of entries within each time slot. The x-axis represents the time slots, and the top-10 ranked motif list of each time bin is compared with the one preceding it and the Kendall's tau coefficient is calculated accordingly. As displayed in Figure 3, the Kendall's tau coefficient slightly varies but is always close to 1 for all time slots. Moreover, the curves also exhibit a very stable trend.



(a) CDR (b) Facebook

Figure 3: Stability of Motifs

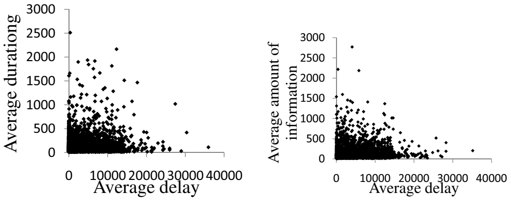
Cost versus motif properties



(a) CDR (b) CDR

Figure 4: Cost vs motif properties

We shall revisit now the star and chain-type motif distribution in the CDR and Facebook datasets shown in Figure 2. The fact that there are more star-type motifs in the Facebook than in the CDR datasets, and more chain-type motifs in the CDR than in the Facebook motifs may be explained by the cost of each interaction. To further study this phenomenon, we compute the correlations between the cost of communication and the properties of the corresponding motifs.



(a) Average duration (b) Amount of info

Figure 5: User profiles vs delay

Figure 4 depicts the correlations between the size and diameters of each motif and its associated cost. The cost of each phone communication is derived from the duration of each call in the CDR dataset, assuming a linear relationship between cost and duration. It can be observed that as the size of motif increases, neither the duration nor the cost associated with the motif increase; on the contrary, for most large motifs there is a very small associated cost.

Speed vs. Amount vs. User Profile

In this section, we analyze the correlations among the speed of information propagation within the motifs, the amount of information being propagated, and the user profiles. Figure 5 (a) depicts the average duration of the calls made by users in the motif *vs.* the speed of information propagation. Note how users that tend to talk for a long time (*e.g.*, with long duration of calls) tend to propagate information faster than users with smaller average call durations. That is, if you talk a lot on the phone, you are likely to propagate the information quicker. In Figure 5 (b) we study the amount of information that is propagated within the motifs *vs.* the speed of information propagation. Here the amount of information that is propagated is derived from the duration of the calls in the motif.

4. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel approach to analyze the patterns of information propagation within communication-based social networks by leveraging the temporal annotations of social communication data and the topological structure. Based on two new concepts: *communication motifs* and *maximum-flow communication motifs*, we have carried out extensive experiments with two real datasets, and have found that the proposed motif structures have desirable features such as temporal stability and are able to capture important aspects of human communication, such as information propagation.

5. REFERENCES

[1] G. Kossinets, J. M. Kleinberg, and D. J. Watts. The structure of information pathways in a social communication network. In *KDD*, 2008.

[2] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *KDD*, 2009.

[3] R. MIlo, S. Shen-Orr, S. Itzkovitz, et. al,. Network motifs: Simple building blocks of complex networks. *Science*, 298(25):824–827, 2002.

[4] J.-P. Onnela, J. Sarama, J. Hyvonen, et. al,. Structure and tie strengths in mobile communication networks. *PNAS*, 104(18):7332–7336, May 2007.