Persistent Cascades: Measuring Fundamental Communication Structure in Social Networks

Steven Morse
Operations Research Center, MIT
Draper Laboratory
Cambridge, MA
stmorse@mit.edu

Marta C. González
Dept. Civil and Environmental Eng.
MIT
Cambridge, MA
martag@mit.edu

Natasha Markuzon
Draper Laboratory
Cambridge, MA
nmarkuzon@draper.com

***Abstract*—We define a new structural property of large-scale communication networks consisting of the persistent patterns of communication among users. We term these patterns "persistent cascades," and claim they represent a strong estimate of actual information spread. Using metrics of inexact tree matching, we group these cascades into classes which we then argue represent the communication structure of a local network. This differs from existing work in that (1) we are focused on recurring patterns among specific users, not abstract motifs (e.g. the prevalence of triangles or other structures in the graph, regardless of user), and (2) we allow for inexact matching (not necessarily isomorphic graphs) to better account for the noisiness of human communication patterns. We find that analysis of these classes of cascades reveals new insights about information spread and the influence of certain users, based on three large mobile phone record datasets. For example, we find distinct groups of weekend vs. workweek spreaders not evident in the standard aggregated network. Finally, we create the communication network induced by these persistent structures, and we show the effect this has on measurements of centrality.**

I. Introduction

A natural question to ask in the study of communication in social networks is: *do social networks exhibit a recurring pattern of information spread?* In this paper we propose methods which indicate the answer may be *yes*. Specifically, we present a method of extracting what appear to be the underlying communication structures from the "noisy" information available in large-scale datasets.

We focus our attention on mobile phone records, also termed call detail records (CDRs), because they provide a unique opportunity to study the large-scale, unfiltered communication patterns of individuals among their friends. Unfortunately, this breadth of knowledge — in time, space, and demographics — comes at the expense of depth, since we have no information about the purpose or content of communication as we might in social media or email records. Our approach attempts to solve this problem by finding persistent patterns that strongly imply meaningful communication is taking place.

A. Related work

A standard approach to translate raw communication data into a meaningful network is to aggregate user activity over some time period T (e.g. a week or month) into a static graph. For example, we can require that a call is reciprocated to consider two users social contacts (and assign them an edge) as in [21], and choose T such that it gives some stable representation (see [11]).

An alternative approach is to include temporal knowledge, an interpretation broadly called *temporal networks* ([9]), which often improves our understanding of structure and community both at an aggregate and individual level. For example, in [19], they observe that the change in a user's frequent contacts over time adheres to an apparent upper bound, or social capacity, that stays relatively constant for a user even as his/her contacts evolve.

The temporal approach seems especially appropriate in the study of information spread, which is by nature causal and time-dependent. Strong properties of human interaction have emerged by including temporal information. One such is the property of "burstiness" — that is, people tend to communicate in short, active bursts followed by long periods of inactivity. The tendency for non-Poissonian, heavy-tailed inter-event communication times has been observed in many contexts (for example, [27] studies email virus propagation, [4] mobile phone communication, and [8] both mobile phone and email), and shown to slow diffusion dynamics ([7], [24]) except under certain conditions ([18]).

A critical question in the study of information spread in temporal networks is determining *what (or if) information is being spread during an observed communication event*: is this call/email/tweet random, social, information-related, etc. In datasets like social media posts or email the answer is usually obvious from the text content; for example, using Twitter hashtags as in [13], [15], [6]. However, in data like CDRs where we only have the metadata of each event, a solution is not obvious. In [1], they contrast the calling patterns immediately following an emergency (bombing, earthquake) with the rest of the call events, and find systematic differences in the timing and spread of information. The implication is that we are more sure "real" information spread is occurring following an emergency, and therefore the contrast of patterns between this spread and what we infer through a standard aggregated approach indicates the latter is an inaccurate estimate.

This type of "cascading" information spread — i.e., a single user initiating a call to a few contacts, who then call several more, and so on — is of great interest in answering our question of the communication event's purpose, since (broadly), a cascade implies non-random, or causal, action (see

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

[3], [6], [13], [15], [25] and others).

We can make an even stronger claim to the meaningfulness of a particular observed communication pattern, if we see the pattern repeated again and again over time. One area of research in this theme is searching for frequent structures, or "motifs," in the temporal network (independent of the specific users involved). In temporal networks this amounts to a recurrent, isomorphic, time-respecting subgraph. For example, [14] analyzes cascade motifs in blog posts and reposts, and [29] analyzes frequency of communication motifs in both CDR and Facebook wall post data. Subsequent works have presented increasingly robust and efficient frameworks for identifying and analyzing these recurrent temporal subgraphs ([2]), often using a comparison to some null model ([10]).

B. Contributions

In this paper, we propose a method for extracting the recurrent patterns of information spread among users in a social network. This extends previous work by incorporating the idea of *similarity* (rather than isomorphism), and by considering user-specific patterns (rather than abstract motifs like a chain or triangle). We call the patterns *persistent cascades*, and claim that they represent a strong estimate of the meaningful, underlying communication structure of their local social network. We show that analysis of the persistent cascades bolsters previous work like burstiness and social capacity, and also leads to new insights like a habitual hierarchy among friends in information spreading and the existence of weekday- or weekend-only tendencies in some communication patterns. We also contrast the centrality of users using a static approach against a network weighted with the persistent cascade structures.

The paper is organized as follows. In Section II, we define a cascade, present two similarity measures to measure persistence, and introduce a *cascade-weighted network* using these structures. In Section III, we analyze the persistent cascades and classes and contrast the centrality of users inferred from a standard aggregated network against the cascade-weighted network. Section IV summarizes our findings and presents avenues for future work.

II. METHODS

Consider an observed pattern where user A calls users B and C, who then call users D, E and F, and then we observe this same pattern, or something similar, repeated every few days or weeks. We term these *persistent cascades*, and claim the pattern leads to two very reasonable assumptions: (1) it is more likely that calls in a persistent cascade indicate meaningful social interactions than calls not observed in one, and (2) it is highly likely that persistent cascades correspond with information spread. We seek to define, find, and analyze these patterns.

A. Defining a cascade

Consider a temporal graph $G = (V, E)$ which represents the communications between users over some large time period $T = [t_{\text{begin}}, t_{\text{end}}]$, such as one month. Let each node $v \in V$ represent a user who participates in some number of communication events during period T, and let each edge $e \in E$ represent a communication event which we encode as a 4-tuple $e_i = (s_i, d_i, t_i, \delta_i)$ consisting of the initiator (s_i), the receiver (d_i), the time of the event (t_i), and its duration (δ_i).

We define a *time-respecting* path as any sequence of edges $(e_1, e_2, ..., e_k)$ such that for any consecutive pair e_i, e_j in the sequence, we have that $d_i = s_j$ and $t_i + \delta_i \leq t_j$. We define a Δt*-connected* path as a time-respecting path such that $t_k - t_1 \leq \Delta t$. From these definitions, one can construct Δt-connected subgraphs that contain some time-respecting subset of *all* the events within Δt (e.g. [10]).

However, in pursuit of understanding information spread patterns, we make an assumption that the information *originates from a single user*, and every user receives the information at the *earliest possible time*. This implies there is a single in-edge to each user, and creates a rooted, directed tree structure. Intuitively, this shifts focus from the structure of the call patterns to the structure of the information spread, since we will only capture the first occurrence of "information" being passed.

Formally, this assumption leads to the construction of a rooted, directed, Δt-connected tree which we term a *cascade*. This term, and its construction, follows closely that in [24].

Denote a cascade with root r as C_r, denote the set of all cascades for root r with maximum time interval Δt and total time period T as $\mathcal{C}_r(T, \Delta t)$, and use superscripts as necessary to distinguish multiple cascades with the same root. For example, we might have the set of all cascades for some root a:

$$\mathcal{C}_a(T = 1 \text{ mo},\ \Delta t = 24 \text{ hrs}) = \left\{C_a^1, C_a^2, C_a^3\right\} \tag{1}$$

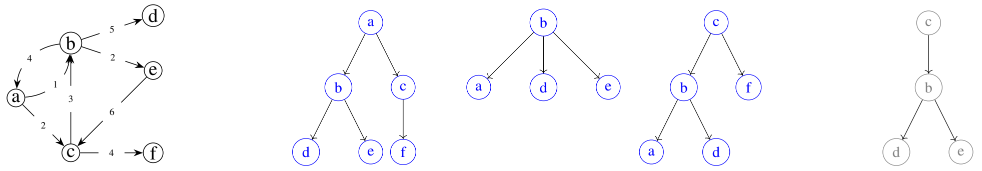
Note we require that the intervals not overlap: i.e. no calls from C_a^1 can also be in C_a^2, etc.

An example of cascade construction from a network with all temporal information is shown in Figure 1

We make two notes about this definition before proceeding. First, notice that for any cascade, its subtrees are also (usually) cascades. For example, in Figure 1, note that the cascade with root a has a subtree corresponding to the cascade with root b. This is by design: we do not know the true information originator, so we should consider each possible "root" user in his or her own right in the analysis of persistence that follows.

Second, consider a root node who is very consistent in the users he calls, but these users are then subsequently very *in*consistent. Then the overall cascades generated from this root will be dissimilar, and therefore ignored in the subsequent analysis. This is again by design: we are concerned with persistent information *spread*, not just consistent calls from a particular user to certain friends. Cascades that are only similar in the first level do not indicate the root is a strong originator of information.

Fig. 1. Simplified illustration of cascade extraction from a temporal graph. For clarity, we examine a network with only 6 nodes. **(a)** Full temporal information ($\Delta t = 6$ units, times depicted on edges). **(b)** Three valid cascades given this temporal snapshot. Note that there is no time ordering of children within a cascade. **(c)** Invalid cascade because: (c-b-e) is not a time-connected path, and missing the edge (c-f).

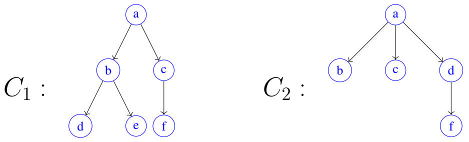


(a) (b) (c)

B. Similarity measures

Measuring similarity of cascades, as defined, is an inexact tree matching problem. We now define two similarity measures, both standard in the literature: normalized tree edit distance, and reach set similarity (measured with a Jaccard index).

a) Tree edit distance: Edit distance is the process of counting the minimum number of insertions, deletions, or mutations required to transform one string into another. One can extend this concept to trees. Denote the tree edit distance between two trees (or cascades) C_1 and C_2 as $\text{TED}(C_1, C_2)$, which maps two cascades to a nonnegative integer. As an example, consider the following two trees :



To change C_1 into C_2, we can delete d and e, mutate c into d, and add c again, giving $\text{TED}(C_1, C_2) = 4$ (note this is the same to change C_2 into C_1).

A canonical algorithm for computing this distance is due to Zhang and Shasha ([28]), which we implement with the `zss` package available at [5]. We can now define a similarity measure using this distance as follows.

Definition 2.1: Tree Edit Distance similarity. Define the normalized tree edit similarity as

$$s_{\text{TED}}(C_1, C_2) \overset{\text{def}}{=} 1 - \frac{2 \cdot \text{TED}(C_1, C_2)}{|C_1| + |C_2| + \text{TED}(C_1, C_2)}. \tag{2}$$

and note s_{TED} lies on $[0, 1]$.

This definition is due to [16], who also prove that the corresponding distance metric $1 - s_{\text{NTED}}$ meets the triangle inequality. Note we make every edit operation unit cost.

Using the example trees above, we now compute $s_{\text{TED}} = 1 - \frac{2 \cdot 4}{6+5+4} = \frac{7}{15} \approx 0.47$.

b) Reach set: Consider the un-ordered set of all nodes in a tree. For a cascade, this corresponds to all users who the root reached during the time period Δt, and potentially received some information. We term this the *reach set* of a cascade (similar to concepts in [15], [23]).

A simple first approximation of the similarity of two cascades is by comparing their reach sets. Let $R(C_i)$ denote the reach set of a cascade C_i. Now, given two cascades C_1 and C_2, define the similarity measure s_{RS} as the Jaccard index of the two reach sets, that is

Definition 2.2: Reach Set similarity. Given two cascades C_1 and C_2, and their reach sets $R(C_1)$ and $R(C_2)$, define

$$s_{\text{RS}}(C_1, C_2) \overset{\text{def}}{=} \frac{|R(C_1) \cap R(C_2)|}{|R(C_1) \cup R(C_2)|}. \tag{3}$$

and note s_{RS} lies on $[0, 1]$.

Continuing with the previous example, we have $s_{\text{RS}}(C_1, C_2) = \frac{5}{6} \approx 0.83$.

C. Persistence

We now would like to group cascades together which all share some minimum pairwise similarity, and so are in a relaxed (but well-defined) sense the "same cascade." This group now represents various incarnations of some fundamental communication structure. We call these groups *persistence classes*, and the elements of each group *persistent cascades*, and they are the main object of our analysis.

Definition 2.3: Persistence class. Define the i-th persistence class of root r, similarity threshold ℓ in time period T over intervals Δt, as the set

$$\mathcal{P}_r^i(\ell, T, \Delta t) = \left\{ C_r^1, C_r^2 \in \mathcal{C}_r(T, \Delta t) \ : \ s_*(C_r^1, C_r^2) \geq \ell \right\} \tag{4}$$

and the collection of all persistence classes for a particular root as $\mathcal{P}_r(\ell, T, \Delta t)$.

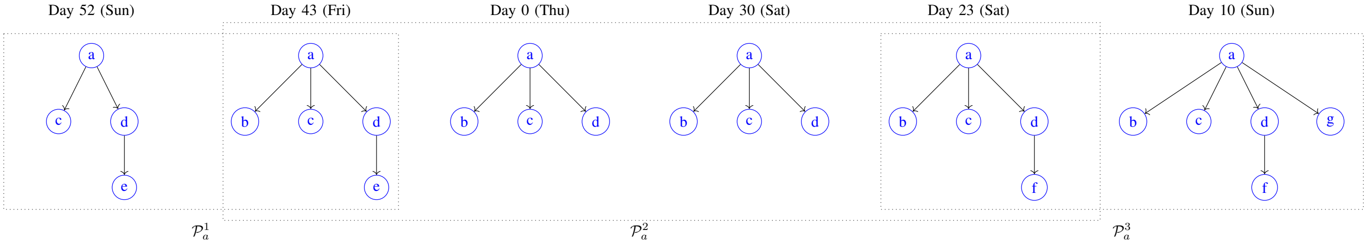
Definition 2.4: Persistent cascade. Define a persistent cascade as any cascade C_r^i such that $C_r^i \in \mathcal{P}_r(\cdot)$, for some r.

Note we may also choose to ignore any persistence classes below a certain size. The minimum size is 2 by construction, but we may decide based on the parameters T and Δt that a minimum size of 3 or more is appropriate.

To find these classes, our definition and Eq. (4) leads us directly to an agglomerative clustering approach with complete-linkage — that is, define the similarity between two clusters U and V as $s(U, V) = \min s_*(U_i, V_j), \forall i \in U, \forall j \in V$ where U_i, V_j represent cascades within U and V. Then the clusters at iteration k, such that every pairwise similarity within the cluster is $\geq s_k$, represent persistence classes with $\ell = s_k$.

However, this assumes that each cascade falls uniquely into one class, which we can imagine is not always true: a spreading pattern among work friends may overlap with the

Fig. 2. Actual set of cascades for a root a over a 60-day period. Six persistent cascades are shown, each from temporal subgraphs with $\Delta t = 24$ hours. Dotted rectangles depict the persistence class groupings. We see a clear set of "core friends" (nodes b, c, d), and slight variations incorporating other groups. We also see the overlap that occurs when a cascade appears to fit in multiple classes. Labeled above each cascade is the day of the week.



pattern among social friends, and there may be cascades that are not clearly in one class or the other.

So we instead adopt a graph-theoretic interpretation of the complete-linkage approach: represent each data point (cascade) as a vertex in a graph $H(s_k)$ such that each any two vertices with similarity $\geq s_k$ are connected. Then the clusters at iteration k correspond to the maximal completely connected subgraphs in H, also known as the maximal cliques. [17]

Now, applying this technique, consider the collection of persistence classes $\mathcal{P}_a$ depicted in Figure 2, taken from City A. Here, we see a core pattern consisting of root a calling b, c, and d, captured in $\mathcal{P}_a^2$. Then, we see two variations on this core structure: $\mathcal{P}_a^1$ which incorporates e, and $\mathcal{P}_a^3$ which incorporates f and g. Since they are mostly weekend calls, we might easily imagine this being a core group of social friends, with variations possibly for family or work acquaintances.

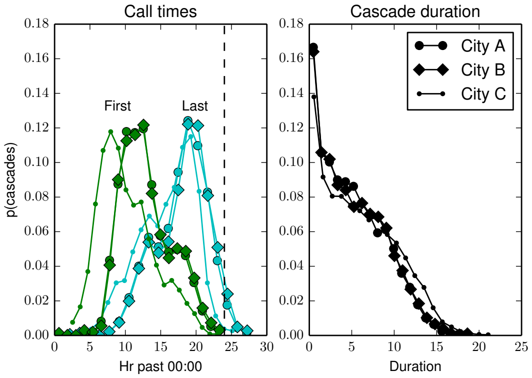
We make two notes on our methodology of identifying persistence. First, we are only doing pairwise comparison between cascades which share a root node, leaving out groupings such as different initiators who disseminate information to the same people. It has the effect of maintaining focus on analysis of the roots, instead of the broader role or persistence of a cascade pattern itself. Second, note that it is conceivable that unrelated call events could happen consistently in the same order among the same people and get picked up mistakenly as persistent. Not knowing the actual content of the calls, we can only say that persistence, as defined, indicates a very high likelihood of information spreading.

III. RESULTS

a) Data: The datasets are CDRs from three cities and their greater metropolitan area: two mid-size European cities (City A and City B) and one Central American city (City C). The data consist of caller, callee, and time stamp for each phone call or SMS event recorded by the carrier. (Location information is also recorded, but not used in this study.)

For a given month, in City A, there are about 280k (280×10^3) unique users, making a total 5.8 million call/SMS events. City B has about 212k unique users making 3.9 million call/SMS events. City C sees about 1.7 million unique users each month, making 154 million call/SMS events. Of all unique users, the fraction who initiate at least 3 cascades to at least 2 other users varies somewhat by region: City A and City B have about 30-35% meeting this criteria (98k and 72k, respectively), while City C has about 65% (1.1 million). This may be due to the City C dataset being more recent, and so there is an overall higher level of mobile phone activity.

Fig. 3. Distribution of call times and duration among persistent cascades. The left plot shows the distribution of times for the *first* (i.e. earliest) and *last* (latest) calls in a cascade. The right plot shows the resulting distribution of total duration of a cascade.



A. Cascade size, time and duration

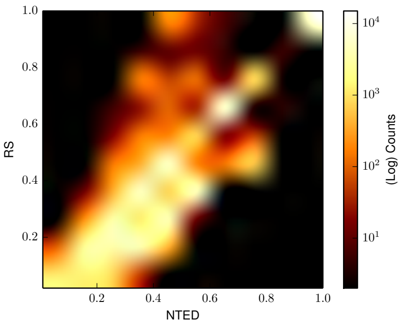
We find that most persistent cascades (e.g. 71% of the sample in City A) are among 3 contacts (the minimum necessary to constitute a cascade). The largest persistent structures involve 20-30 people (for example, in City A, we find a persistent class with cascades of 37-39 users, but note persistent cascades with more than 6 people constitute less than 1% of the sample).

Figure 3 shows the distribution of first and last call times in a (persistent) cascade, and the resulting distribution of cascade durations. This was done on a random sample of 10^4 root users in all 3 cities over a period of 1 month. The call times follow the expected workday pattern of a morning peak around 9-10 a.m., and another peak before nightfall around 8 p.m.

We also see from the right plot in the figure that most persistent cascades are very short — usually everyone is called within an hour — which echoes earlier work on the burstiness of communication. There is also a large group of cascades with durations between 5-10 hours, suggesting information spread is either very rapid, or unfolding over a morning or afternoon, but rarely lasting all day.

This evident short attention span in the cascades led us to avoid analysis of longer time periods (48, 72 hours or longer). Longer time periods also may decrease the possibility of the cascade representing information spread. It may be fruitful to consider a shorter interval, such as 12 hours, to attempt to capture morning vs. evening cascading action (e.g. work vs. social), or a sliding window approach. We leave exploration to future work.

Fig. 4. Heat map depicting the correlation between NTED and RS metrics on a sample of 5×10^4 pairs of cascades with the same root (over approximately 10^4 different roots). The number of pairs where $s_{\text{RS}}(x, y) = 1.0$ but $s_{\text{TED}}(x, y) < 1.0$ is surprisingly small — only about 0.5% of the sample — suggesting that cascades among the same users tend to occur in the same order. Note: the colors are log-scaled for visualization.



B. Similarity measure correlation and habitual hierarchy

We now examine the relationship between the two similarity measures introduced in Section II: tree edit distance (TED) and reach set (RS). Based on a random sample of 5×10^4 pairs of cascades from City B, the measures have a Pearson correlation coefficient of $\rho = 0.91$. (Results are similar in other datasets.)

It is possibly surprising that the correlation is so high. For example, consider the group of cascades with s_{RS} of 1.0 and s_{TED} less than 1.0, and note that this group represents less than 0.5% of the sample. This shows that when two cascades involve the same people, they nearly always involve them in *the same order*. (And if not, we would see more pairs with dissimilar structure (low TED) but similar reached users (high RS).) This observation suggests there is a *habitual hierarchy* of information spread among social contacts.

Note: the main performance bottleneck in computing all persistence classes for a particular dataset is the TED measure. However, the correlation between measures shows RS is a close approximation in most cases. It is also much easier to compute; so, if computing $\mathcal{P}_*$ under both measures, one can compute RS similarity first, and only compute TED similarity as necessary for s_{RS} above some low threshold. Finally, since we are only considering classes with the same root, the clustering step is parallelizable. Using these speedups, we could build all persistence classes for a single city, with both similarity measures and $T = 1$, in about 30 minutes.

C. Tendency for weekday vs. weekend information spread

Consider the set of all cascades (not necessarily persistent) that a given (root) user initiates in the course of some period T, for example a month. Since most active users tend to make some calls every day, we might expect these cascades to be evenly distributed over each day of the week.

In Table I we examine all cascade initiators in each city with at least one persistent class and at least 3 persistent cascades. If we consider all cascades of this group (not just persistent ones), we see that there is an even mix throughout the week, as expected: nearly all users are generating cascades (that is, making calls to multiple people) on some mix of both weekend and weekdays. Very few users ($< 1\%$) are active exclusively on weekdays and/or weekends.

TABLE I
DISTRIBUTION OF ROOT NODES BY TIME OF CASCADE: PERSISTENT CASCADES REVEAL A TENDENCY FOR WEEKEND OR WEEKDAY INFORMATION SPREADING

Cascade type	Dataset	Only Weekend	Mix	Only Weekday
All	City A	<1%	99.2%	<1%
	City B	<1%	99.4%	<1%
	City C	<1%	99.8%	<1%
Persistent	City A	1.8%	82.5%	15.6%
	City B	2.6%	83.8%	12.9%
	City C	2.5%	84.2%	13.3%

Note. "Only" weekend/weekday signifies at least 90% of events. Fridays designated as the weekend.

However, if we examine only *persistent* cascades, two new groups emerge: a large portion of root users who only initiate persistent cascades on weekdays, and a slightly smaller portion who only initiate on weekends. These two extremes constitute over 15% of all root users, while the same extremes measured in all cascades are $< 1\%$. This is a complement to the observation that people have different mobility similarities to weekend and weekday contacts, in [26].

In other words, for these two groups, although they make calls throughout the week, their role in spreading information appears to be specialized: their only persistent patterns of information spread happen during either weekday (i.e., work week) hours or weekend hours, but not both. Their other communication is sporadic, or random, and one might easily conclude, not meaningful.

D. Long-term persistence

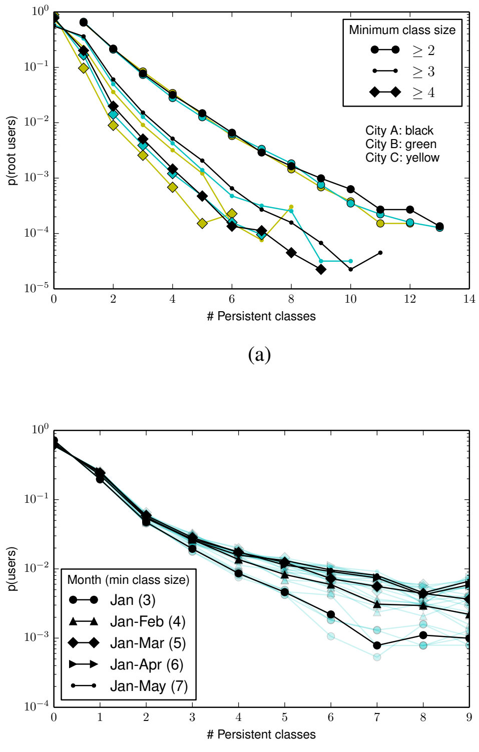
Now we turn our attention to observations of the persistent structures over longer periods of time ($T > 1$ month). One intuitive property we expect to see emerge is the idea of *long-term persistence*. Specifically, if the persistent classes represent the fundamental underlying communication structure of the network, we expect them to persist over long periods of time — that is, user's should continue to generate cascades which "fit" into existing classes.

First, in Figure 5(a), note the decline in the distribution of persistent classes as we increase the minimum size requirement (i.e., for a user a, enforce that $|\mathcal{P}_a^i(\cdot)| \geq k$, for all i, and increase $k = 2, 3, 4, ...$). This is an expected effect of increasing requirements within a finite time. For a minimum size of 4 cascades, only about a tenth of the population has even one persistent class.

If there were no long-term persistence of these classes, then we would see no class growth over time, and the distributions of persistent classes would decline as we increase their minimum size requirement, regardless of the time period.

However, in Figure 5(b), the opposite happens. As we increase the time period and the minimum size requirement, the distribution of persistent classes increases somewhat and stays generally the same, especially for the 90% of the population with 3 or fewer classes. This implies that our intuition is correct, and many (if not most) of the persistent classes continue to grow as time goes on.

We can also be more precise by checking, for example, how many specific users with 1 persistent class after 1 month, still have 1 persistent class after 2, 3, 4, and 5 months, etc. We find that about 65% of users with a single persistent class (of size ≥ 5) after 3 months of observation, will still have a single persistent class (now of size ≥ 6) after 4 months of observation. And about 71% with a single class after 4 months will again have a single class (now of size ≥ 7) after 5 months. This is remarkable consistency, and suggests a strong predictability of calling habits.



(a)

(b)

Fig. 5. Distributions of the proportion of users by their # of persistent classes. In (a) we increase the requirement for persistence but fix T. In (b) we increase the requirement for "persistence" from 3 to 7 cascades in the class as we increase T from 1 to 5 months. (Excludes the top 1% of users.) Note that (a) demonstrates an expected finite-time effect, but in (b) the distributions are nearly identical, especially for users with 0-2 classes (who constitute over 90% of the sample), suggesting long-term persistence and bounded social capacity. (The black plot depicts the average over 5 samples of 5×10^3 random users; samples depicted in light green.)

E. Cascade-weighted network

Now consider applying this knowledge of persistent structure back to a static structure, and observing the effect on, in particular, centrality. Specifically, for a network $G = (V, E)$, weight the subset of edges E_C that are present in at least one persistent cascade with $w_c = \alpha \in [0.5, 1]$ and all $e_n \in E \setminus E_C$ with $w_n = 1 - \alpha$. Now with $\alpha = 0.5$ we recover the standard aggregated network, and with $\alpha > 0.5$ we are putting extra weight on the "persistent" edges which we claim carry more meaning.

This results in a network of about 278k nodes and 505k edges, with about 45k users having at least one persistent class of 2 or more cascades (counts are for City A). Setting α in $[0.5, 1)$, we find a Large Connected Component (LCC) comprising 80-85% of the total network for all three datasets (cf. [21]). With $\alpha = 1$, the LCC splits into several thousand smaller subgraphs, the largest usually being about 2k nodes. This echoes previous results that show the inability of information to reach any sort of macroscopic diffusion when traveling solely through information cascades [24], and one could consider it another version of the general result of slowed diffusion in temporal networks [7].

We now consider the weighted degree (or node strength [20]) of a user i, defined $k_i = \sum_j A_{ij}$, where A is the adjacency matrix of G and $A_{ij} = w_c$ if $(i, j) \in E_C$, w_n if $(i, j) \in E \setminus E_C$, and 0 otherwise. We examine a 1-month time period in City A, for both the unweighted (i.e., $\alpha = 0.5$) and cascade-weighted ($\alpha > 0.5$) networks. We use the s_{TED} measure for this analysis, with $\ell = 0.8$. We observe the effects of the weighting in Table II, which presents the overlap of central and non-central users for both networks as measured by degree, when $\alpha = 0.5$ against when $\alpha = 0.9$.

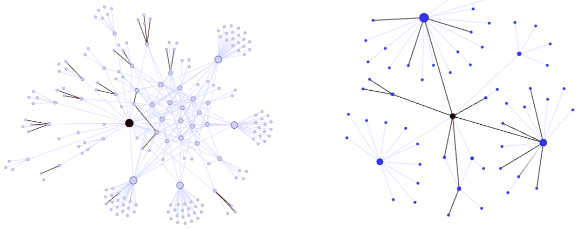
We note several groups that emerge: first, the large group of users (about 7% of the total users) that are only central in the cascade-weighted network. This suggests a group of users with unremarkable importance as measured in a naïve way by counting calls, but who play a pivotal role in the persistent communication patterns of their social network. Similarly, a large group of influential users in the standard unweighted network disappears when we begin weighting cascades, implying their centrality was only due to a web of edges corresponding to mostly random calls. And lastly, we note that a large portion of the network has their status essentially unchanged.

TABLE II
CONTRAST OF TOP RANKED USERS (BY DEGREE) IN THE STANDARD UNWEIGHTED VS. CASCADE-WEIGHTED NETWORK. USERS IN **BOLD** (6.6% OF TOTAL POP.) ARE HIGHLY CENTRAL IN INFORMATION SPREAD, BUT ARE UNNOTICED USING A STANDARD APPROACH.

		Weighted	
	k_i (degree) rank	*Bottom ranked*	*Top ranked*
Unweighted	*Bottom ranked*	195,248 (83.9%)	**15,357 (6.6%)**
	Top ranked	18,020 (7.7%)	10,261 (4.4%)

* Bottom rank = lower 90% of users, top rank = top 10% of users

Fig. 6. Example central users in the (left) unweighted network and (right) cascade-weighted network. User of interest depicted as black nodes, all others as blue. Edges present in a persistent cascade depicted in black.



IV. CONCLUSION AND FUTURE WORK

In this paper we introduced a novel way of estimating the real communication structure of a social network, called persistent cascades, using methods of inexact tree matching and agglomerative clustering. This approach extends existing work by keeping focus on individuals (instead of motifs) and allowing a relaxed sense of similarity (instead of isomorphism). We showed that these persistent structures tend to be "bursty" and follow circadian patterns, in line with previous work. We observed that the high correlation of structural- and user-centric metrics implied a habitual hierarchy in communication. We discovered a tendency for day-of-the-week-specific information spreading among about 25% of the population that is completely hidden using standard methods. We demonstrated the long-term persistence of these structures, over a period of several months, and the bounded social capacity this implies. Finally, we introduced a cascade-weighted network and revealed a group of about 6-7% of the population that is highly central in persistent communication, and thus stand-outs in the weighted network, but insignificant using a standard approach.

We expect there is potential in coupling these insights of communication structure with the knowledge of *mobility* that we get with many datasets; for example, do we find high similarity of mobility patterns [26] of users within most classes? Do information spreaders exert observable influence on their social contacts' movement habits? We also hope to examine the effect the cascades have under a diffusion model.

In conclusion, we hope this paper contributes a new method of understanding the persistent patterns of human communication in large-scale networks, and that in future work we may be able to extend this to a deeper understanding of the dynamics of centrality and information spread in communication networks in the urban space.

ACKNOWLEDGMENT

This work was partially supported by internal Draper Laboratory funding, the Gates Foundation and the UN foundation. S.M. thanks R. Di Clemente for many valuable discussions. We thank GranData for a data collaboration agreement.

REFERENCES

[1] J. Bagrow, D. Wang, A.-L. Barabási. "Collective response of human populations to large-scale emergencies," PLoS ONE, vol. 6 (3), Mar. 2011.

[2] P. Bogdanov, M. Mongiovi, A. Singh. "Mining heavy subgraphs in time-evolving networks," IEEE Intl. Conf. on Data Mining, 2011.

[3] J. Borge-Holthoffer, R. Baños, S. González-Bailón, Y. Moreno, E. Estrada (ed.). "Cascading behavior in complex socio-technical networks," J. Complex Networks, vol. 1, Mar. 2013.

[4] J. Candia, M. González, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabási. "Uncovering individual and collective human dynamics from mobile phone records." J. Phys. A: Math. and Theor., vol. 41, 2008.

[5] T. Henderson, S. Johnson. Python implementation of Zhang-Shasha tree edit distance algorithm. https://github.com/timtadh/zhang-shasha.

[6] C. Hui, Y. Tyshchuk, W. Wallace, M. Magdon-Ismail, M. Goldberg. "Information cascades in social media in response to a crisis: a preliminary model and case study," WWW Conf., SWDM, 2012.

[7] J. Iribarren and E. Moro. "Impact of human activity patterns on the dynamics of information diffusion," Phys. Rev. Letters, vol. 109, 2009.

[8] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, J. Saramäki. "Small but slow world: how network topology and burstiness slow down spreading," Phys. Rev. E., vol. 83, 2010.

[9] D. Kempe, J. Kleinberg, and A. Kumar. "Connectivity and inference problems for temporal networks," STOC, 2000.

[10] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, J. Saramäki. "Temporal motifs in time-dependent networks," J. Stat. Mech., 2011.

[11] G. Krings, M. Karsai, S. Bernhardsson, V. Blondel, J. Saramäki. "Effects of time window size and placement on the structure of an aggregated communication network," EPJ Data Science, Springer, 2012.

[12] M. Lahiri and T. Berger-Wolf. "Structure prediction in temporal networks using frequent subgraphs," IEEE Symp. on Comp. Intell. and Data Mining (CIDM), 2007.

[13] K. Lerman, R. Ghosh. "Information contagion: an empirical study of the spread of news on Digg and Twitter social networks," Proc. AIII Conf. on Weblogs and Social Media, 2010.

[14] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst. "Patterns of Cascading Behavior in Large Blog Graphs," Proc. of the 2007 SIAM Conf. on Data Mining, 2007.

[15] C.-T. Li, Y.-J. Lin, M.-Y. Yeh. "The Roles of Network Communities in Social Information Diffusion," IEEE Intl. Conf on Big Data, 2015.

[16] Y. Li and C. Zhang. "A metric normalization of tree edit distance," Front. Comp. Sci. China, vol. 5 (1), 2011.

[17] C. Manning, P. Raghavan, H. Schütze. Introduction to information retrieval. NY: Cambridge Univ. Press, 2008, ch. 17, pp. 378-85.

[18] G. Miritello, E. Moro, R. Lara. "Dynamical strength of social ties in information spreading," Phys Rev E vol. 83, 2011.

[19] G. Miritello, R. Lara, M. Cebrian, and E. Moro. "Limited communication capacity unveils strategies for human interaction," Scientific Reports, vol. 3, 2013.

[20] M. E. J. Newman. "Analysis of weighted networks," Phys. Rev. E, vol. 70, 2004.

[21] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. de Menezes, K. Kaski, A.-L. Barabási, J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. New J. of Phys (9). 2007.

[22] G. Palla, A.-L. Barabási, T. Vicsek. "Quantifying social group evolution," Nature, vol. 446, 2007.

[23] R. Pan, J. Saramäki. "Path lengths, correlations, and centrality in temporal networks," Phys Rev E, vol. 84, 2011.

[24] F. Peruani, L. Tabourier. "Directedness of information flow in mobile phone communication networks," PLoS ONE, vol. 6 (12), 2011.

[25] L. Tabourier, A. Stoica, F. Peruani. "How to detect causality effects on large dynamical communication networks: a case study," IEEE, 2012.

[26] J. Toole, C. Herrera-Yaquë, C. Schneider, M. González. "Coupling human mobility and social ties," J. R. Soc. Interface, vol. 12, Feb. 2015.

[27] A. Vazquez, B. Rácz, A. Lukács, A.-L. Barabási. "Impact of non-Poissonian activity patterns on spreading processes," Phys Rev Letters, vol. 98, 2007.

[28] K. Zhang, D. Shasha. "Simple fast algorithms for the editing distance between trees and related problems," Siam J. Comput., vol. 18 (6), 1989.

[29] Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, W.-C. Lee. "Communication motifs: a tool to characterize social communications," CIKM, 2010.