

Persistent Cascades

Measuring patterns of information spread

Steven Morse, Marta González, and Natasha Markuzon

December 7, 2016

Operations Research Center, MIT
Human Mobility and Networks Lab, MIT
Draper Laboratory, Cambridge

IEEE Big Data 2016



Central question

Do social contacts have recurring, temporal patterns of spreading information?

Central question

Do social contacts have recurring, temporal patterns of spreading information?

We claim certain persistent group conversation patterns we term *persistent cascades* give indications the answer is **yes**. In this talk, we

- Define persistent cascades.
- Propose a method for extracting these structures from “raw” communication metadata.
- Give analysis of these structures which reveals new properties of human communication (and confirm known ones).

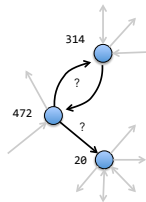
The Problem

Given communication metadata (i.e. no knowledge of content), how do we determine which events represent “meaningful” interactions?

Approaches we might try...

- Edge-weighting
- Temporal motifs
- Frequent subgraphs
- Learning probabilistic structure

Caller	Callee	Time
472	314	3970084
30	407	3970085
772	896	3970088
703	705	3970088
...
314	472	3982191
27	33	3982392
472	20	3983092
...
314	472	3411193
425	282	3411194
101	99	3411296
233	881	3413297



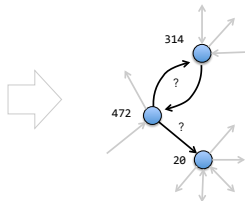
The Problem

Given communication metadata (i.e. no knowledge of content), how do we determine which events represent “meaningful” interactions?

Approaches we might try...

- Edge-weighting
- Temporal motifs
- Frequent subgraphs
- Learning probabilistic structure

Caller	Callee	Time
472	314	3970084
30	407	3970085
772	896	3970088
703	705	3970088
...
314	472	3982191
27	33	3982392
472	20	3983092
...
314	472	3411193
425	282	3411194
101	99	3411296
233	881	3413297



What if we see person A call person B, who calls person C and D, and then the same pattern (or *something similar*) repeats again on a regular basis?

Defining persistent cascades

Defining a cascade

Construction: Select a (root) node, and recursively add each of the root's callees, who they called, etc. through some interval Δt . Keep earliest event in case of multiple calls. Repeat for all (disjoint) intervals in a window T .

Cascades (Collection)

Denote the collection of cascades with root r over time window T in intervals Δt as

$$\mathcal{C}_r(T, \Delta t) = \{C_r^i\}$$

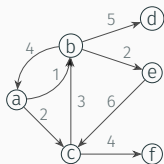
Defining a cascade

Construction: Select a (root) node, and recursively add each of the root's callees, who they called, etc. through some interval Δt . Keep earliest event in case of multiple calls. Repeat for all (disjoint) intervals in a window T .

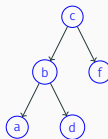
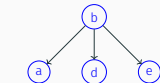
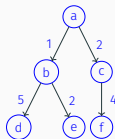
Cascades (Collection)

Denote the collection of cascades with root r over time window T in intervals Δt as

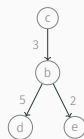
$$\mathcal{C}_r(T, \Delta t) = \{C_r^i\}$$



Full network



Valid cascades



Not valid

(Normalized) Tree Edit Distance

Let $TED(\cdot, \cdot) : C \times C \rightarrow \mathbb{N}$ denote the *tree edit distance* between two cascades [8]. Then, for any two cascades C^1, C^2 , define

$$S_{NTED}(C^1, C^2) \stackrel{\text{def}}{=} 1 - \frac{2 \cdot TED(C^1, C^2)}{|C^1| + |C^2| + TED(C^1, C^2)}$$

Note $S_{NTED} \in [0, 1]$. (See [3].)

Measuring similarity – Tree Edit Distance

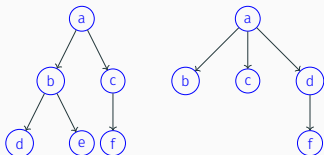
(Normalized) Tree Edit Distance

Let $TED(\cdot, \cdot) : C \times C \rightarrow \mathbb{N}$ denote the *tree edit distance* between two cascades [8]. Then, for any two cascades C^1, C^2 , define

$$S_{NTED}(C^1, C^2) \stackrel{\text{def}}{=} 1 - \frac{2 \cdot TED(C^1, C^2)}{|C^1| + |C^2| + TED(C^1, C^2)}$$

Note $S_{NTED} \in [0, 1]$. (See [3].)

Example:



$$S_{NTED} = 1 - \frac{2 \cdot 4}{6 + 5 + 4} = \frac{7}{15} \\ \approx 0.47$$

Reach Set similarity

Let $R(C^i)$ represent the unordered list of nodes in a cascade C^i , its *reach set*. Then, for two cascades C^1 and C^2 , define

$$s_{RS}(C^1, C^2) \stackrel{\text{def}}{=} \frac{|R(C^1) \cap R(C^2)|}{|R(C^1) \cup R(C^2)|}$$

(i.e., the Jaccard index of the two cascades' reach sets.)

Measuring similarity — Reach Sets

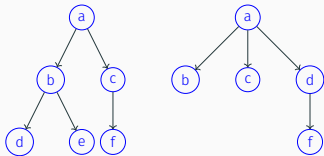
Reach Set similarity

Let $R(C^i)$ represent the unordered list of nodes in a cascade C^i , its *reach set*. Then, for two cascades C^1 and C^2 , define

$$s_{RS}(C^1, C^2) \stackrel{\text{def}}{=} \frac{|R(C^1) \cap R(C^2)|}{|R(C^1) \cup R(C^2)|}$$

(i.e., the Jaccard index of the two cascades' reach sets.)

Example:



$$s_{RS} = \frac{5}{6} \\ \approx 0.83$$

Persistence Class

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

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

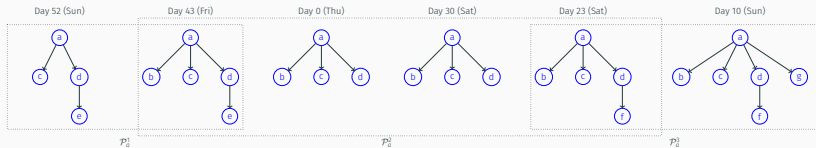
Finding persistence

Persistence Class

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

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

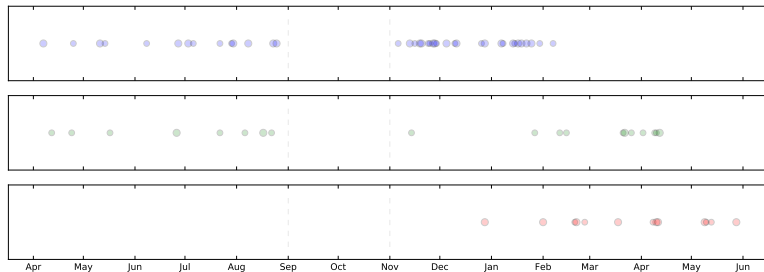
Example:



Findings

Examples: long term persistence

- In the figure below, we see three examples of long term persistent classes, with each dot representing the occurrence of a persistent cascade in the class.
- In the first subplot, there appears to be a crescendo of activity followed by the class vanishing (possibly event planning?).
- In the last subplot, we see a group form (possibly due to holidays?).



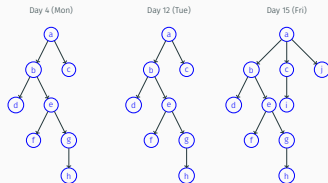
Dots represent the occurrence of a persistent cascade. Size of the dot represents the number of participants in the cascade. Each subplot corresponds to a single persistence class. (The subplots are unrelated.) Dashed lines indicate missing data.

Examples: large structures

We typically only find 3- and 4-node trees that are persistent at the time scale of a year. However, we find surprisingly large persistent structure at the scale of months.

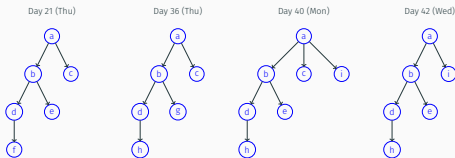
10 distinct users,

1 month:



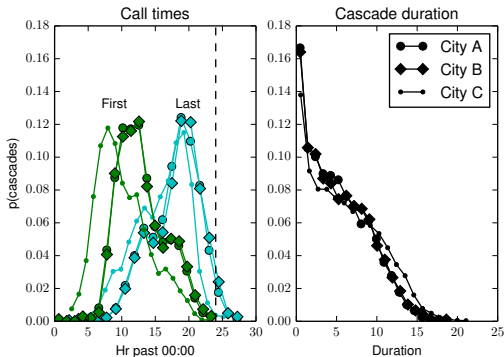
9 distinct users,

2 months:



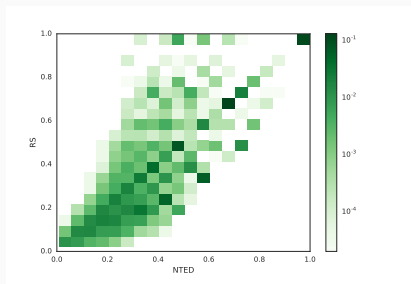
Short duration

- The first/last call in persistent cascades follow normal circadian rhythms.
- However the total cascade duration is usually < 2 hours.
- This gives indication of an urgency (or **burstiness**) in persistent communication.



Correlation of similarity measures

- Comparing the NTED and RS similarity measures on a sample of 5×10^4 pairs of cascades, we find they are highly correlated ($\rho = 0.91$). There is less correlation for intermediate values.
- **Habitual hierarchy.** We find very few examples ($< 1\%$ of sample) of cascades with the same users but different structure (i.e. $RS = 1.0$, $NTED < 1.0$). This indicates that when the same individuals are communicating, it is almost always *in the same order*.



Cascades reveal new roles in information spreading

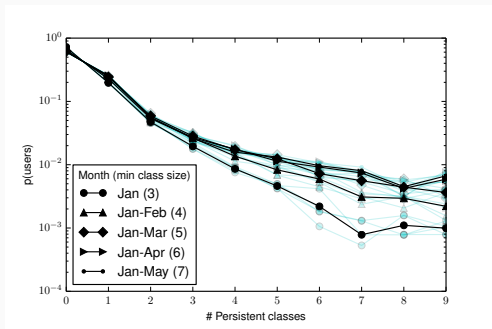
- Consider the cascades (not necessarily persistent) that an individual generates in a month. They are, for $> 99\%$ of the population, a mix of weekends and weekdays.
- However, if we consider only *persistent* cascades, two new groups emerge: one that is only active in persistent communication on exclusively weekends, and another on weekdays.

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%

- Fridays included as weekends. "Only" signifies $> 90\%$ of events

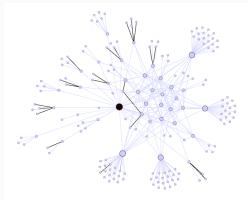
Long-term persistence and predictability

- We expect a persistence class to grow over time, as the group continues to generate new incarnations of the same pattern.
- This implies **predictability**. For example, out of individuals with one class after 3 months of observation, about 65% will still have the same (though now larger) single class after 4 months.

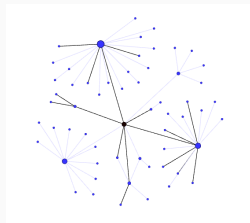


Persistence reveals new central individuals

- Weighting edges according to their membership in a persistent cascade reveals new groups of “central” individuals in the network.
- The two subgraphs below correspond to a snowball sample around high centrality users (top 10%) using unweighted or weighted degree centrality. (Persistent edges in black.)
- In (a), we see a highly connected node, but whose connections are non-persistent. In (b), we see a only moderately connected node, but who is involved in several large persistent group conversations and therefore highly central in the weighted network.



(a) Unweighted



(b) Weighted

Wrapping up

Conclusion and future work

- **Effect on spreading dynamics.** We now have a subset of interactions that is non-Poissonian *and* carries more weight — how much effect do these persistent group conversations have on information/opinion spread?
- **Prediction.** How well are these structures modeled by, for example, a non-homogeneous point process, and can we use such a model to predict future calls? What do the estimated parameters tell us about the nature of the group and their conversation?
- **Relationship with mobility.** Do we find high similarity of mobility patterns of users within most classes? Do information spreaders also exert observable influence on their social contacts' movement habits?

Questions?

References I



A.-L. Barabasi.

The origin of bursts and heavy tails in human dynamics.

Nature, 435, 2005.



L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki.

Temporal motifs in time-dependent networks
Temporal motifs in time-dependent networks
Temporal motifs in time-dependent networks.

J. Stat. Mech, 2011.



Y. Li and C. Zhang.

A metric normalization of tree edit distance.







J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. De Menezes, K. Kaski, A. L. Barabási, and J. Kertész.

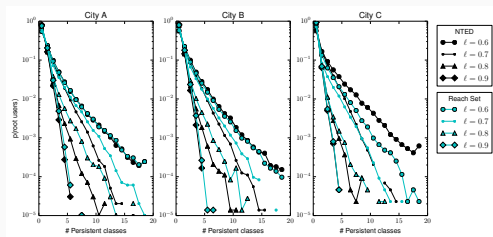
Analysis of a large-scale weighted network of one-to-one human communication.

New Journal of Physics, 2007.

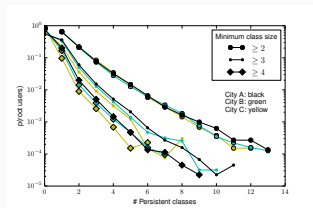
References II

-  F. Peruani and L. Tabourier.
Directedness of Information Flow in Mobile Phone Communication Networks.
2011.
-  J. L. Toole, C. Herrera-yaqu, C. M. Schneider, and M. C. Gonza.
Coupling human mobility and social ties.
Journal of the Royal Society, Interface, 12:1–14, 2015.
-  A. Vazquez, B. Rácz, A. Lukács, and A. L. Barabási.
Impact of non-poissonian activity patterns on spreading processes.
Physical Review Letters, 2007.
-  K. Zhang, R. Statman, and D. Shasha.
On the editing distance between unordered labeled trees.
Information Processing Letters, 42(3):133–139, 1992.

Sensitivity



(a)



(b)

Centrality comparison

Table 1: 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		<i>Bottom ranked</i>	<i>Top ranked</i>
Unweighted	<i>Bottom ranked</i>	195,248 (83.9%)	15,357 (6.6%)
	<i>Top ranked</i>	18,020 (7.7%)	10,261 (4.4%)

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