UNIVERSITY OF TWENTE.

Network analysis for trend prediction in social media

Nelly Litvak,

University of Twente,
Stochastic Operations Research group

Joint work with Yana Volkovich, Anna Tolkacheva, Marijn ten Thij

Budapest, May 8, 2014

Introduction

- Trending topics on Twitter in relation with different real-life events such as elections, social protest, or sports events
- Can we provide informative measures that characterize the difference between trends?
  - Dynamic dependencies
  - Connected components
- Project funded by Google Faculty Research Awards
- Agenda
  - Experimental setting, definitions
  - Experiments
  - Modeling and analysis
  - Connected components

Datasets

- **Source**
  Social network - Twitter
- **Method**
  Collect all tweets which contain particular word for some periods of time

Datasets

- **Key words**
  - **Maidan** (rus)
  - **Euromaidan** (ukr)
  - **Sochi olympics** 2014 (rus)
  - **Putin** (rus)
  - **Berkin Elvan alive** (turk)
  - Some other words for short periods

Datasets

- **Time periods**
  - **Maidan**
    from 16-11-2013 till 02-1-2014
  - **Euromaidan**
    from 02-12-2013 till 09-3-2014
  - **Olympics**
    from 07-12-2013 till 09-3-2014
  - **Putin**
    from 09-11-2013 till 17-3-2014
  - **Berkin Elvan alive**
    from 07-03-2014 till 11-3-2014

    Some periods have missing days

Datasets

- **Maidan**
  286.984 tweets, 120.996 retweets, 87.498 users
- **Euromaidan**
  2.433.517 tweets, 1.788.604 retweets, 162.582 users
- **Olympics**
  735.849 tweets, 289.269 retweets, 250.569 users
- **Putin**
  879.711 tweets, 333.250 retweets, 227.320 users
- **Berkin Elvan**
  1.856.387 tweets, 1.261.590 retweets, 582.861 users

Definitions

- T the total length of the tracking period
- $t_1, \ldots, t_m$ – subsequent subperiods (e.g. length of one day)
- $G_i = (V_i, E_i)$ – retweet graph period t_i
- V_i – users that tweeted or received a retweet on t_i
- $E_i = \{(u, v) : u \text{ retweeted } v \text{ on } t_i\}$
- $G = \cup_i G_i$

Centrality measures

- In-degrees
  - $D_i(v)$ – in-degree of v in G_i
- Harmonic centrality (Boldi&Vigna, 2013):
  - $d_i(w, v)$ – the length of a directed path from w to v in G_i
  - *Harmonic centrality* $H(v)$ of node $v \in V_i$ is defined as a sum of inverse graph distances from w to v over all $w \in V_i$:

$$H_i(v) = \sum_{w \in V_i} \frac{1}{d_i(v, w)}.$$

- Centralities are computed for each G_i and for G.

Dynamic dependencies

- Let $|V| = n$ be the total number of users in a data base.
- We consider vectors of length n that contain degrees or harmonic centrality scores of each user in a given day or in the complete retweet graph
- We compute correlations between these vectors
  - Between main graph and a graph in each given day
  - Between every 2 graphs of the consequent days
- Correlation measures:
  - Cosine similarity
  - Spearman correlation

Cosine similarity measure

- V – all users that ever tweeted on the topic
- For two vectors $(X(v))_{v \in V}$ and $(Y(v))_{v \in V}$, we define the cosine similarity measure as follows:

$$cos(X, Y) = \frac{\sum_{v \in V} X(v) Y(v)}{\sqrt{\sum_{v \in V} X^2(v)} \sqrt{\sum_{v \in V} Y^2(v)}}. \qquad (1)$$

- The cosine similarity measure for non-negative vectors takes values between 0 (no similarity) and 1 (similarity up to a factor)
- Elements in (1) also define the Pearson's correlation coefficient, and indeed the two measures are closely related (Lee et al. 1988)

Spearman's rho

- Arrange the values of $(X(v))_{v \in V}$ and $(Y(v))_{v \in V}$ in decreasing order
- Let $R_X(v)$ and $R_Y(v)$ be the rank (position) of, respectively, $X(v)$ and $Y(v)$.
- Since the data has many ties, we consider two versions of Spearman's ρ:
  - *Average*: all tied values receive the same, average, rank.
  - *Random*: each tied value receives a unique rank, the order is defined at random.
- Two ways of resolving ties is that the average rank remains $(|V|+1)/2$.

Spearman's rho

- Let $|V| = n$
- The Spearman's rho:

$$\rho(X, Y) = \frac{\sum_{v \in V} R_X(v) R_Y(v) - (n+1)^2/4}{n\sigma(X)\sigma(Y)}, \qquad (2)$$

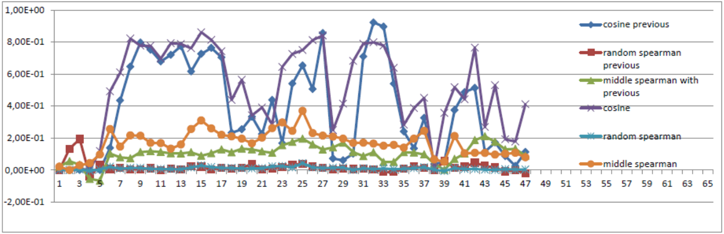
where for $Z = X, Y$

$$\sigma(Z) = \sqrt{\frac{1}{n} \sum_{v \in V} R_Z^2(v) - (n+1)^2/4}.$$

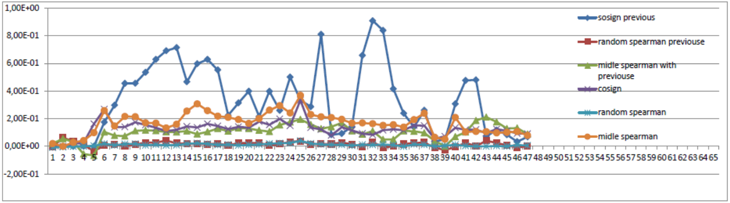
- The difference between average and random way of resolving ties is only in denominator.
- Randomly resolved ties: $\sigma(X) = \sigma(Y) = (n^2 - 1)/12$.
- With average resolution of ties, the values of σ become smaller and this leads to a higher value of ρ. This is quantified exactly (L&vdHoorn 2014).

Experiments

Maidan Degree

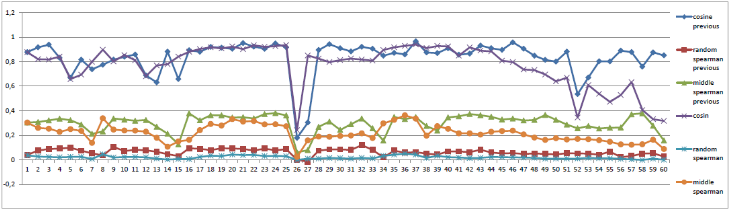


Maidan Harmonic Centrality

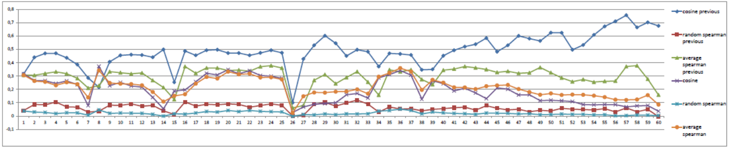


Experiments

Euromaidan Degree

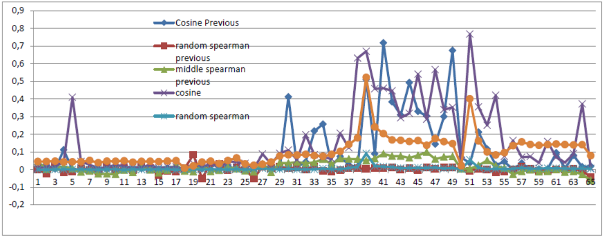


Euromaidan Harmonic Centrality

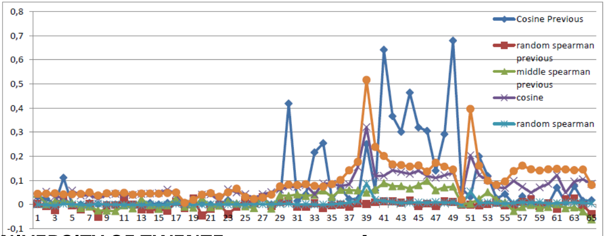


Experiments

Sochi Olympics Degree

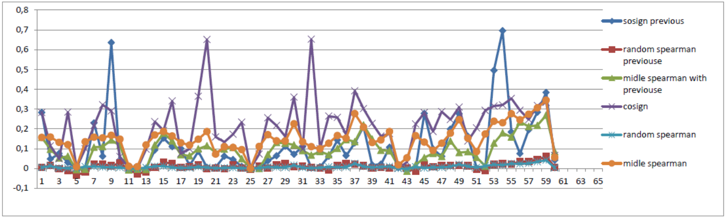


Sochi Olympics Harmonic Centrality

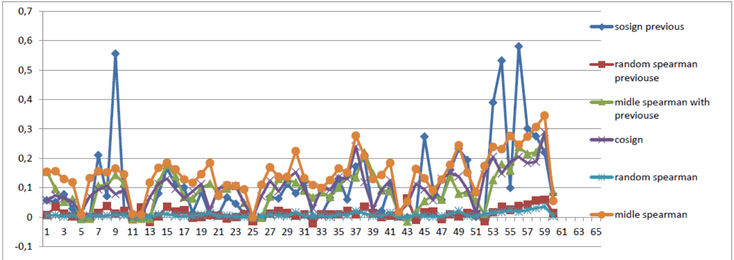


Experiments

Putin Degree

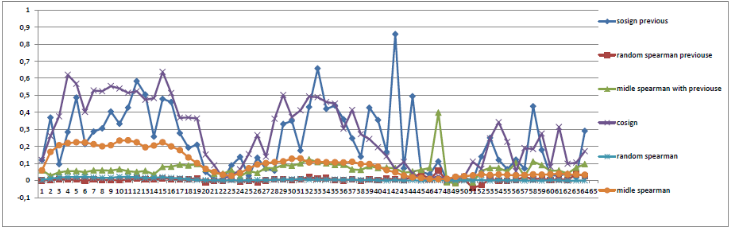


Putin Harmonic Centrality

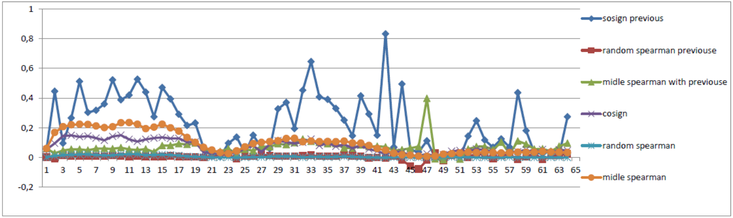


Experiments

'Berkin Elvan Alive' Degree



'Berkin Elvan Alive' Harmonic Centrality



‘Missing’ data

- Important feature of the data is that only a fraction of users in V is present in V_i
- Many tied values of centralities are simply zero’s
- This explains the large difference between random and average resolution of ties for Spearman’s rho
- We model this by assuming that a user tweets on period t_i with probability p_i

The model

- Let $X_i(v)$ be a centrality score of user v in graph G_i
- Multiplicative model:

$$X_i(v) = \begin{cases} \alpha_i(v)U(v), & \text{w.p. } p_i; \\ 0 & \text{w.p. } 1 - p_i. \end{cases} \tag{3}$$

- $U(v)$ popularity of user v,
- $\alpha_i(v)$ shows how this popularity scales in time period t_i with respect to centrality score X.
- We assume that $(\alpha_i(v))_{i \geqslant 1}$ are i.i.d.
- $(U(v))_{v \in V}$ i.i.d. random variables with regularly varying (power law) distribution U:

$$P(U > x) = L(x)x^{-\gamma}, \quad x > 0, \gamma > 1. \tag{4}$$

Here $L(x)$ is a slowly varying function, that is, $\lim_{x \to \infty} L(tx)/L(x) = 1$ for all $t > 0$.

Analysis of cosine measure

$$cos(X_i, X_{i+1}) = \frac{\sum_{v \in V_i \cap V_{i+1}} U^2(v)\alpha_i(v)\alpha_{i+1}(v)}{\sqrt{\sum_{v \in V_i} U^2(v)\alpha_i^2(v)}\sqrt{\sum_{v \in V_{i+1}} U^2(v)\alpha_{i+1}^2(v)}}$$

- U^2 is a regularly varying random variable with index $\gamma/2$
- Assuming that for some $\varepsilon > 0$ we have $E(\alpha^{\gamma+\varepsilon}) < \infty$
- It follows from Breiman's theorem (Breiman 1965) that $\alpha_i^2 U^2$ and $\alpha_i \alpha_{i+1} U^2$ are also regularly varying with index $\gamma/2$.
- According to the law of large numbers, as $|V| \to \infty$, we have $|V_i|/|V| \to p_i$ a.s., and by the independence assumption of the time periods, $|V_i \cap V_{i+1}|/|V| \to p_i p_{i+1}$ a.s.

Stability of cosine measure. Case 1.

- In our model

$$cos(X_i, X_{i+1}) = \frac{\sum_{v \in V_i \cap V_{i+1}} U^2(v)\alpha_i(v)\alpha_{i+1}(v)}{\sqrt{\sum_{v \in V_i} U^2(v)\alpha_i^2(v)}\sqrt{\sum_{v \in V_{i+1}} U^2(v)\alpha_{i+1}^2(v)}} \qquad (5)$$

- **Case 1:** $Var(U) < \infty$.
- $\gamma/2 > 1$, then $E(\alpha_i^2 U^2) < \infty$ and $E(\alpha_i \alpha_{i+1} U^2 < \infty$.
- Dividing the nominator and denominator in (5) by $|V| = n$ and letting $n \to \infty$ we obtain

$$\lim_{n \to \infty} cos(X_i, X_{i+1}) = \frac{E(\alpha_i \alpha_{i+1})\sqrt{p_i p_{i+1}}}{\sqrt{E(\alpha_i^2)}\sqrt{E(\alpha_{i+1}^2)}}, \ a.s.$$

Stability of cosine measure. Case 2.

$$cos(X_i, X_{i+1}) = \frac{\sum_{v \in V_i \cap V_{i+1}} U^2(v)\alpha_i(v)\alpha_{i+1}(v)}{\sqrt{\sum_{v \in V_i} U^2(v)\alpha_i^2(v)}\sqrt{\sum_{v \in V_{i+1}} U^2(v)\alpha_{i+1}^2(v)}}$$

- **Case 2:** $E(U^2) = \infty$.
- $\gamma/2 < 1$, the sums in (5) scale roughly as the number of summands to the power $2/\gamma$.
- Classical convergence to stable laws (Gnedenko&Kolmogorov 1968). As $n \to \infty$:

$$cos(X_i, X_{i+1}) \xrightarrow{d} \frac{Z_1(p_i p_{i+1})^{1/\gamma}}{\sqrt{Z_1' + Z_2}\sqrt{Z_1'' + Z_3}}, \qquad (6)$$

- Z_1, Z_1' and Z_1'' are dependent stable $\gamma/2$ random variables
- Z_2 and Z_3 are independent stable $\gamma/2$ random variables
- Positive density on $[0, 1]$.

Stability of Spearman's rho

- Spearman's rho converges to a correct population value
- Let p_i be a probability that a user tweets or receives a retweet on time period t_i
- If p_i is small then, under the assumption that the users tweet independently, $\rho(X_i, X_{i+1})$ is very close to zero, has close-to-normal distribution and small variance
- The expectation of $\rho(X_i, X_{i+1})$ increases when p_i increases
- $\rho(X_i, X_{i+1})$ shows positive dependency if:
  - There is a persistent group of active users, or
  - Users are independent, but a high fraction of users is active each day.
- Work in progress

Predictions using connected components in retweet graph

with Marijn ten Thij, TNO

Predictions using connected components in retweet graph

with Marijn ten Thij, TNO

- Connection between graph structures and important trends

Predictions using connected components in retweet graph

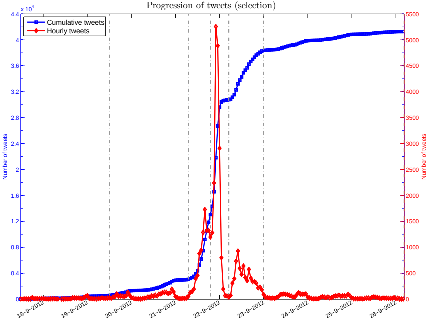
with Marijn ten Thij, TNO

- Connection between graph structures and important trends
- Data: Project X Haren, 21-09-2012

Predictions using connected components in retweet graph

with Marijn ten Thij, TNO

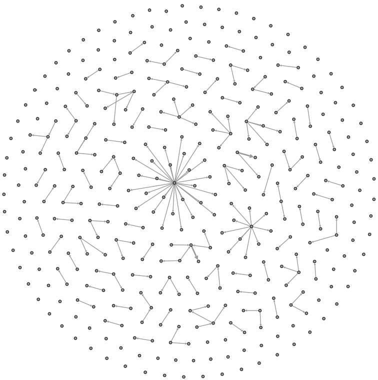
- Connection between graph structures and important trends
- Data: Project X Haren, 21-09-2012



- Undirected retweet graph: a link between two users if one of them retweeted the other

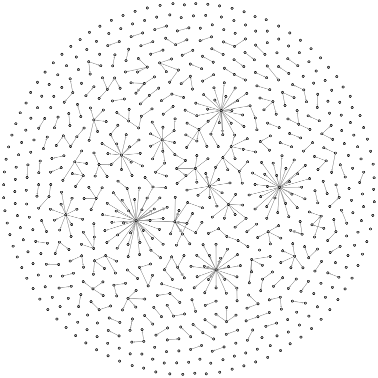
Retweet graph

19-9-2012 12:00



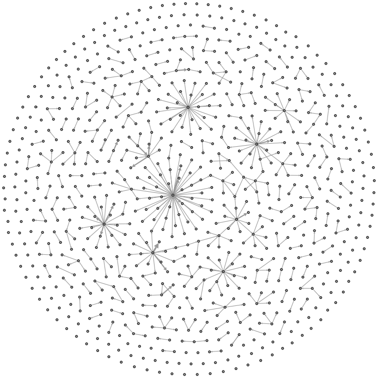
Retweet graph

19-9-2012 23:00



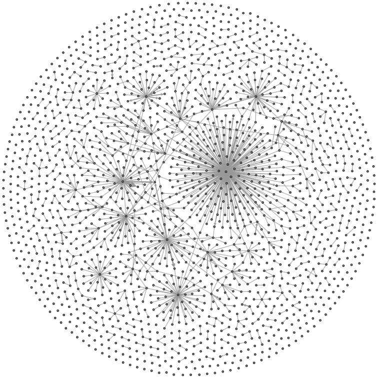
Retweet graph

20-9-2012 00:00



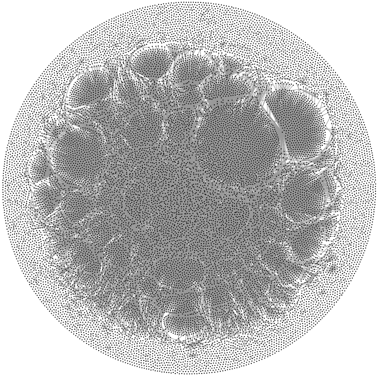
Retweet graph

21-9-2012 07:00

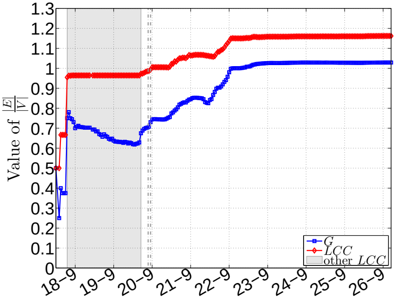


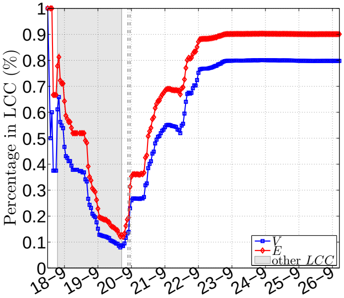
Retweet graph

22-9-2012 05:00



Edge density and largest connected component





Thank you!