**Data**

<http://snap.stanford.edu/infopath/> - 500Mb, information cascade

readings

28 января 2016 г.

20:55

**Data**

<http://snap.stanford.edu/infopath/> - 500Mb, information cascade

[Topic community tracking](http://download.springer.com/static/pdf/535/chp%253A10.1007%252F978-3-642-29253-8_41.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-642-29253-8_41&token2=exp=1454040884~acl=%2Fstatic%2Fpdf%2F535%2Fchp%25253A10.1007%25252F978-3-642-29253-8_41.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Fchapter%252F10.1007%252F978-3-642-29253-8_41*~hmac=a63957b2e0a44eba6b2914600869536777e950e41a6226aa64722f57db39e45f)

Markov random walk

personal affinity (relevance) of user to topic of discussion predicts his chance of engagement.

[Info-Cluster Based Regional Influence Analysis](http://download.springer.com/static/pdf/678/chp%253A10.1007%252F978-3-642-20847-8_8.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-642-20847-8_8&token2=exp=1454041633~acl=%2Fstatic%2Fpdf%2F678%2Fchp%25253A10.1007%25252F978-3-642-20847-8_8.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Fchapter%252F10.1007%252F978-3-642-20847-8_8*~hmac=55e0c09a387e05e57b0bf0955333fdd6a4fdd8eb76fa3034e44b46835e161df3)

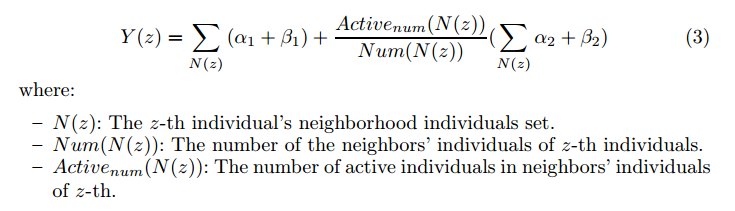
k-means to find location clusters, info clusters - uses both community and location clusters

– Location clustering: It aims to find the clusters (Slc) of locations through some spatial clustering algorithm.

– Community detection: Community detection in social network aims to find groups (Scom) of vertices within which connections are dense, but between which connections are sparse.

modularity function

watt's model - the more people around you are contagious, the more likely you are to get contagious.



nodes get "contagious" or "activated"

[A Fast Approximation for Influence Maximization in Large Social Networks](http://delivery.acm.org/10.1145/2590000/2580063/p1157-lee.pdf?ip=128.138.65.218&id=2580063&acc=ACTIVE%20SERVICE&key=B63ACEF81C6334F5%2E1FC7ACB276C876CF%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=748621414&CFTOKEN=30731620&__acm__=1454042502_21eb5b91cc7f91646d75034655861ea9) (2014)

remove unimportant edges - to speed things up and justify the assumption that not everything has to be connected

"The first obstacle is the expensive cost in calculating the influence spread of a seed set, and the second obstacle is a large number of users in a social network. In this paper, we focus on exploiting the ~~2-hop influence~~ spread of a seed set to overcome the two obstacles."

"an item is generally diffused from a seed within a very small number of hops in online social network services"

"53% favorite markings were propagated through social cascade"

"For example, [1] shows an observation that if a photo is uploaded in Flickr, more than 81% of users who participate in its diffusion are within 2-hops away from a seed."

It means that even if we consider only users who are within 2-hops away from seeds to estimate the influence spread of the seeds, the estimated influence spread of the seeds is experimentally expected to be at least 81% of the exact influence spread. Therefore, exploiting the 2-hop influence spread is sufficiently valid to estimate the influence spread effectively

[1] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In Proceedings of the 18th international conference on World wide web, WWW ’09, pages 721–730, 2009.

"Influence maximization, which is one of famous research problems related to viral marketing"

. Kempe et al. prove that influence maximization under the IC model is NP-hard [8].

"Independent Cascade (IC) model which is one of famous information diffusion models."

user *u* gets a chance to influence its neighbors only once at time *t* with some probability *p*

usually approximated by Monte-Carlo simulations (10,000), expected number of influenced users

hop - modeled by exponential (35% - within a week, median - 60 days, mean - 140 days)

R(distance) - how removed you are from the seed. The more removed, the less people you tell.

[Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph](https://hal.inria.fr/hal-00948889/document)

Network creation:

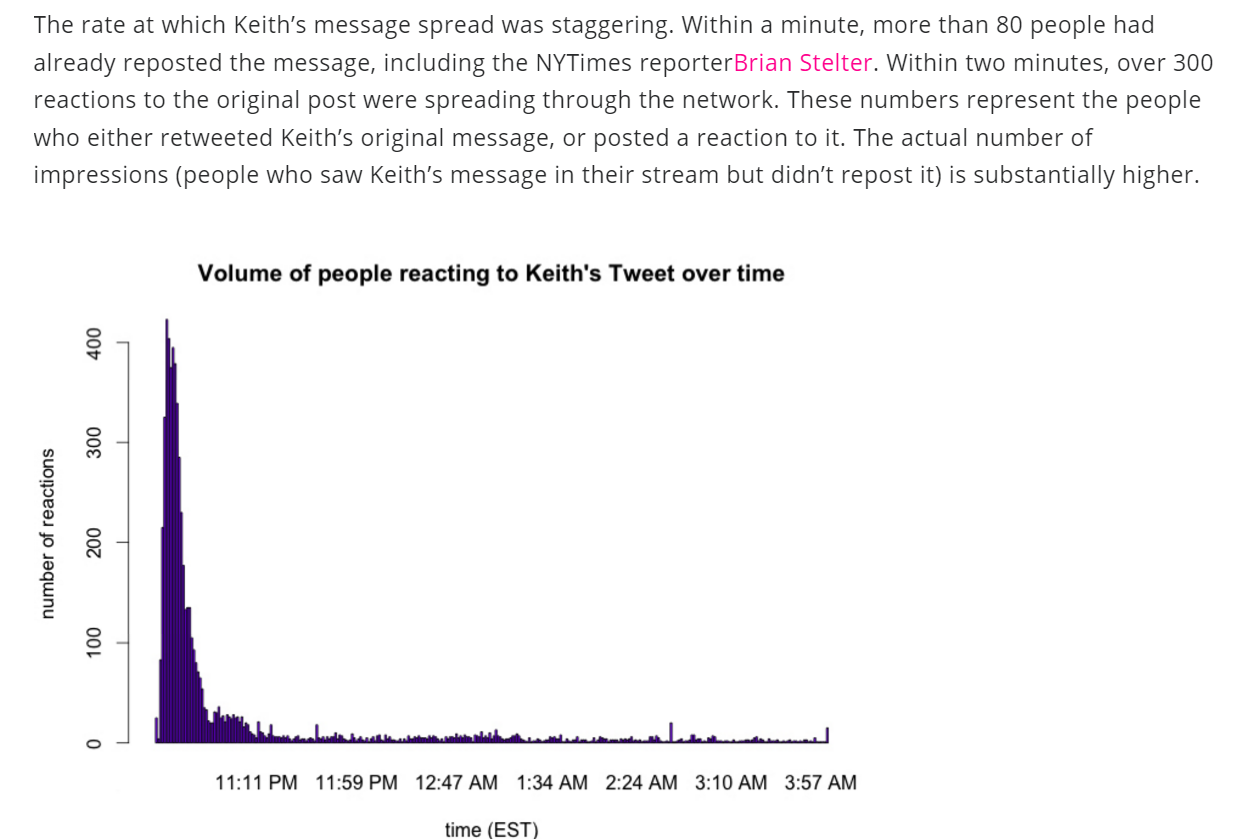
Therefore, we start by identifying all the strongly connected components (SCCs) **[Tarjan algorithm [23]]** that are components with a directed path between any two nodes. In such components, the information can freely circulate, so we abstract each of these components by a single node. After this stage, we obtain a directed acyclic graph (DAG) that is half of the size of the original graph (in terms of number of nodes), still too large to be analyzed. Consequently, the next stage is to group nodes in this DAG based on their connectivity to the largest SCC

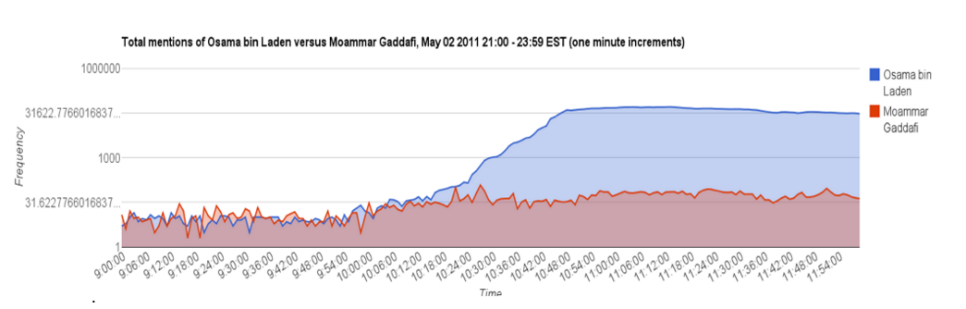
**Paper**

Assumptions

cities are well-mixed, strongly connected components of a network => can be conmbined into one node with parameters of size and etc.

users identify the internet as the most reliable source of information over television and radio (at 3.6 on average, with 5 being totally reliable). [source](http://theconversation.com/hard-evidence-how-does-false-information-spread-online-25567)





[source](http://www.socialflow.com/breaking-bin-laden-visualizing-the-power-of-a-single/)

Therefore, the way information propagates on Twitter is close to how information propagates in real life. Indeed, real life communications are characterized by a high asymmetry between information producers (such as media, celebrities, etc.) and content consumers. Consequently, understanding how information propagates on Twitter has implications beyond computer science.

In particular, we show that regular, abandoned, and malicious accounts are not uniformly spread on the components of the macroscopic structure of the Twitter social graph. This result is important to understand how Twitter is used, where users with a specific usage are, and how to sample Twitter without a significant bias.

Therefore, we start by identifying all the strongly connected components (SCCs) that are components with a directed path between any two nodes. In such components, the information can freely circulate, so we abstract each of these components by a single node. After this stage, we obtain a directed acyclic graph (DAG) that is half of the size of the original graph (in terms of number of nodes), still too large to be analyzed. Consequently, the next stage is to group nodes in this DAG based on their connectivity to the largest SCC

[source](https://hal.inria.fr/hal-00948889/document)

Maximizing the profit of viral marketing is studied as an algorithmic problem by Domingos et al. [4] with a social network modeled as a Markov random field.

CELF++, PMIA, and IRIE?