Diffusion Patterns Of Social Network Posts

Gubanov Alexander, Pu Ida, Mundrievskaya Yuliya

Presenter: Gubanov Alexander
Tomsk State University, derzhiarbuz@yandex.ru
Gubanov Alexander Valerievich, MSc

Analyst in Centre of Applied Big Data Analysis, Tomsk State University, Russia.

Applied mathematician with experience in commercial software development, university teaching (math, programming), inventing (mechatronics and silicate technology) and military service (mandatory).

Current fields of interest - complex network analysis in sociology and construction 3D printing.
Centre of Applied Big Data Analysis is a team of specialists in technic (programmers, mathematicians) and humanities (sociologists, psychologists). We use statistics, data science and network science approaches to analyse social and psychological data. Some of our projects:

- Studying interconnection between student’s digital trace and educational achievements
- Studying the affect of social media content on real life wellbeing of human
- Studying the structure of extremist, self-harm and suicidal behavior in social network
...
**Vkontakte** is the most popular Russian social network with open API and chronological newsfeed ordering.

**Acutors** - users and groups.

**Links** - friendships and subscriptions.

The **unit of information** - post on the wall of user or group, that is visible in **newsfeeds** of friends and subscribers.

1. The wall
2. Author’s post
3. Repost
What we want to know about spreading?

- **Why?** Is the information viral or does it need social support?
- **What?** The information: is it spreads because it’s so effective or it was just started to spread from the best place?
- **Who?** Which actors are make the greatest impact in information spreading?

**Cascade** is a sequence of pairs (actor_id, publication_time):

- (38761, 15:00)
- (4598, 15:07)
- (988713, 15:31)
- ...

**Node**: an actor

**Time**: moment of repost

**Black edge**: social tie

**Green arc**: Possible transmission through tie

**Cascade** combined with social network gives pattern of spreading
OBJECT FEATURES

What’s challenging with our object?

- The **local** (city or region) information cascades are relatively small (50-400 reposts)
- Social network is **partially relevant**
- Channels of spreading are **partially observable**

That’s why it’s hard to use many existing approaches
On the left, we can see:

1. Significant nodes candidates
2. Outsiders (nodes that have a lot of friends but nobody reposts from them)
3. Clusters of users

On the right, we can see that pattern graphs are quite different for commercials and for memes.
VISUAL APPROACH

- Pattern of spreading for multiple cascades

We can see

1. Significant (often used) paths of information spreading
2. Information brokers
3. Active, but non significant users
All diffusion patterns are result of some stochastic process

**Diffusion model** for each link between infected \( i \) and suspicious \( j \) defines the such function \( P_{ij}(t) \) that at time interval \((t, t+dt)\) information spreads through this link with probability \( P_{ij}(t)dt \).

For example, for classic infection SI process \( P_{ij}(t) \) is a **constant** (so-called infection rate \( \theta \))

\[
\frac{dP_j(t)}{dt} = \theta N_j(t)
\]

SI process, where \( P_j(t) \) is a probability for actor \( j \) to be infected at moment \( t \). \( N_j(t) \) - number of \( j \)'s neighbors, infected at this moment.
DIFFUSION MODEL

\( \theta \) - the infection rate (\textit{virulence}) through observable tie \((\theta \geq 0)\)

\( \rho \) - the infection rate through unobservable tie (\textit{background}) \((\rho \geq 0)\)

\( \kappa \) - \textit{conformism} (social pressure coefficient) \((-1 \leq \kappa \leq 1)\)

\( \delta \) - \textit{decay} (information obsolescence coefficient) \((\delta \geq 0)\)

\[
\frac{dP_j(t)}{dt} = \rho \sum_{i \in A(t)} e^{-\delta(t-t_i)} + \tau_\kappa(j, t) \theta \sum_{i \in A_j(t)} e^{-\delta(t-t_i)}
\]
Preliminary results

Estimations are consistent on the set of simulated cascades.

### Parameter Estimation

- $\theta = 0.2$
- $\rho = 0.001$
- $\delta = 0.2$

Model network for testing convergence. Green node is a source of infection.
Infection rate and decay are correlated (obviously)

Heatmap of likelihood function
Red dot - true parameters
Green cross - MLE

θ/δ is the probability of spreading through an edge at least once during infinite time. This value tends to have better estimation.

But it is possible to separate θ from δ on the set of cascades (δ is cascade-independent).
**PRELIMINARY RESULTS**

**conformism** is very fuzzy. We keep it away from the model for now.

**background** is separated from **infection rate** surprisingly well.

Heatmap of **likelihood** function

- **Red dot** - true parameters
- **Green cross** - MLE

Heatmap of **likelihood** function

- **Red dot** - true parameters
- **Green cross** - MLE
PRELIMINARY RESULTS

the **same** source and the **same** theme

$\theta = 1.6e-7$
$\rho = 4.2e-11$
$\delta = 2.3e-4$

$N=93$

$\theta = 4.7e-7$
$\rho = 2.8e-11$
$\delta = 2.4e-4$

$N=124$

$\delta$ is about the same for both cascades. $\rho$ is **1.5** times higher for **red**. $\theta$ is **3** times higher for **green**.

Seems that people like to **spread bad** news but they also like to **manifest good** things.
FURTHER DEVELOPMENT

1. Complete the testing
2. Compare results with existing models
3. Implement parallel calculations and easy-to-use python library
4. Implement the opinion leader detection
ACKNOWLEDGEMENT

1. PATTERNS 2020 Organizers & Logistics for organising the conference especially at this difficult time

2. Jacqueline Daykin for helpful discussion

3. The anonymous reviewers for valuable comments and financial supports for attending the conference

4. Tomsk State University for the visiting scholarship to Goldsmiths University of London