Information Transfer in Social Media

Greg Ver Steeg and Aram Galstyan
USC Information Sciences Institute,Marina del Rey, CA 90292

October 13, 2011

Presentation by Ksenia Kiseleva
Department of Applied Mathematics and Informatics
National Research University
Higher School of Economics

Plan of presentation

- Approaches to characterizing influence
- Using transfer entropy: what is the difference from previous studies?
- Notations and definitions of transfer entropy
- Sampling problems and solutions
- Experiments on synthetic data
- Experiments with Twitter dataset

Recent approaches to influence measuring

- Methods based on explicit causal knowledge
  - Pagerank score
  - Passivity score
  - Using the size of cascade trees
- Algorithms working without knowing the relationship structure
  - Transfer entropy

Notations

$S_X = \{t_j : 0 < t_1 < t_2 \ldots\}$ - timing of tweets

$$B_X(a,b) \equiv \begin{cases} 1 & \text{if } \exists t_j \in S_X \cap (b,a], \\ 0 & \text{otherwise.} \end{cases}$$ - binned variable (I)

$\delta \in \mathcal{R}$ - bin width; T - considered time period

$$P(B_X(t,t-\delta) = X_t) \equiv \frac{1}{T-\delta} \int_{\delta}^{T} dt [B_i(t,t-\delta) = X_t]$$ - probabilities over the binned variables (II)

$$P(B_X(t,t-\delta_0) = X_t, B_X(t-\delta_0, t-\delta_0-\delta_1) = X_{t-1}, \ldots)$$ - joint probability distribution (III)

$$P(X_t^{(t-k)}), \quad where \quad X_t^{(t-k)} = \{X_t, X_{t-1}, \ldots, X_{t-k}\}$$ - simplified written form of joint probability distribution (IV)

Definition of Transfer Entropy

$$H(A|B) = -\sum_{A,B} P(A,B) \log P(A|B)$$ - conditional entropy (I)

$$T_{X \to Y} = H(Y_t | Y_{t-1}^{(t-k)}) - H(Y_t | Y_{t-1}^{(t-k)}, X_{t-1}^{(t-l)})$$ - information transfer (II)

↑ level of uncertainty with knowing Y's history of activity

↑ level of uncertainty with knowing Y's and X's history of activity

Sampling problems and their solutions

- Absence of sufficient data ⟶ To set a minimal level of activity
- Heavy tail in the distribution of the response times ⟶ To set the bins of equal size
- Bias connected with «binned» approach to data modelling ⟶ To use a class of binless entropy estimators

Experiments with synthetic data

Model I

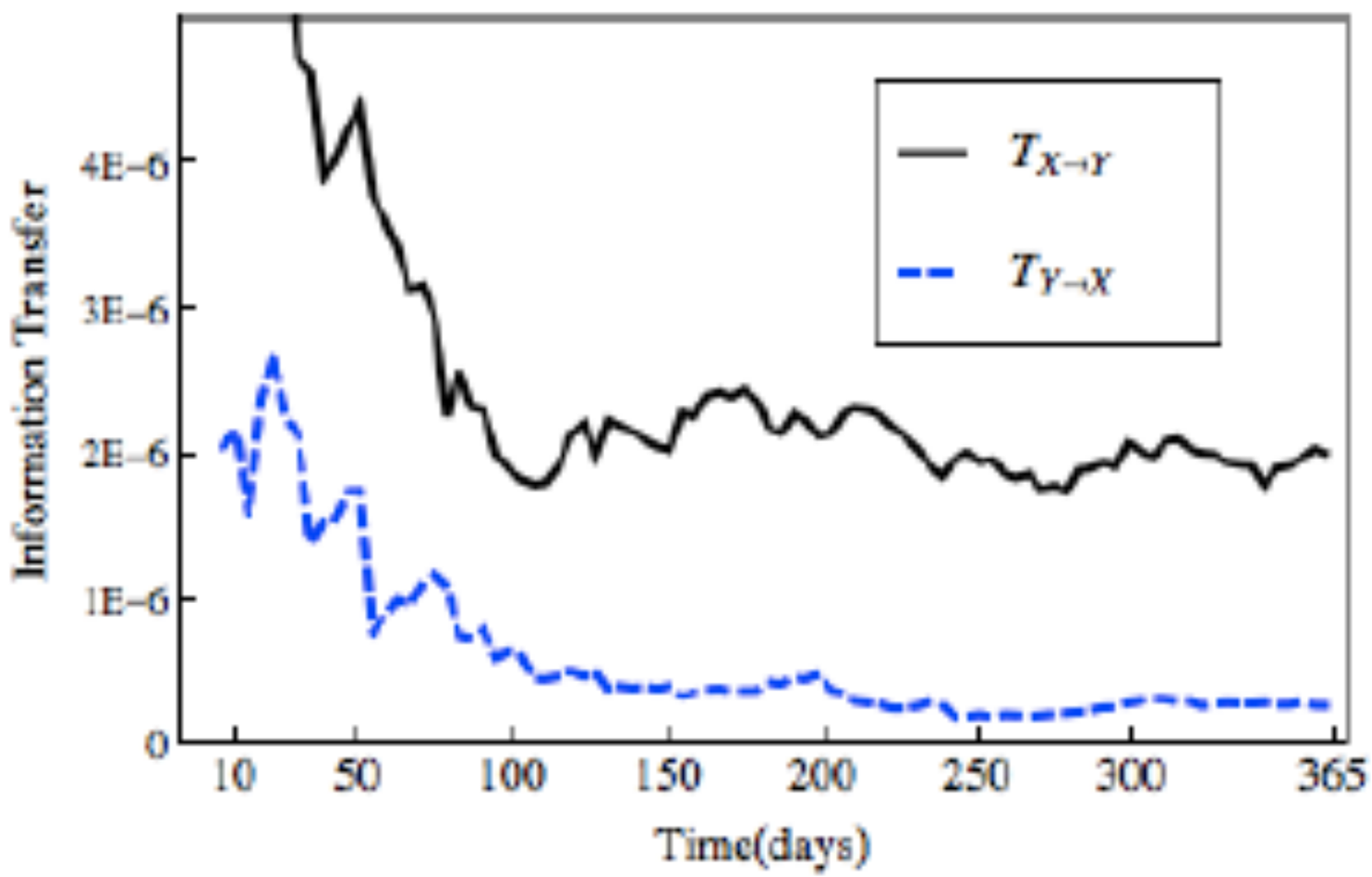
$\lambda_Y(t|S_X^t) = \mu + \gamma \sum_{t_i \in S_X^t} g(t - t_i)$ - Poisson distributed user activity

$g(\Delta t) = \min(1, \left(\frac{1\ \text{hour}}{\Delta t}\right)^3)$ - time dependence of the influence

γ - the strength of influence of X

μ - constant rate of background activity

Transfer Entropy for data generated according to Model 1



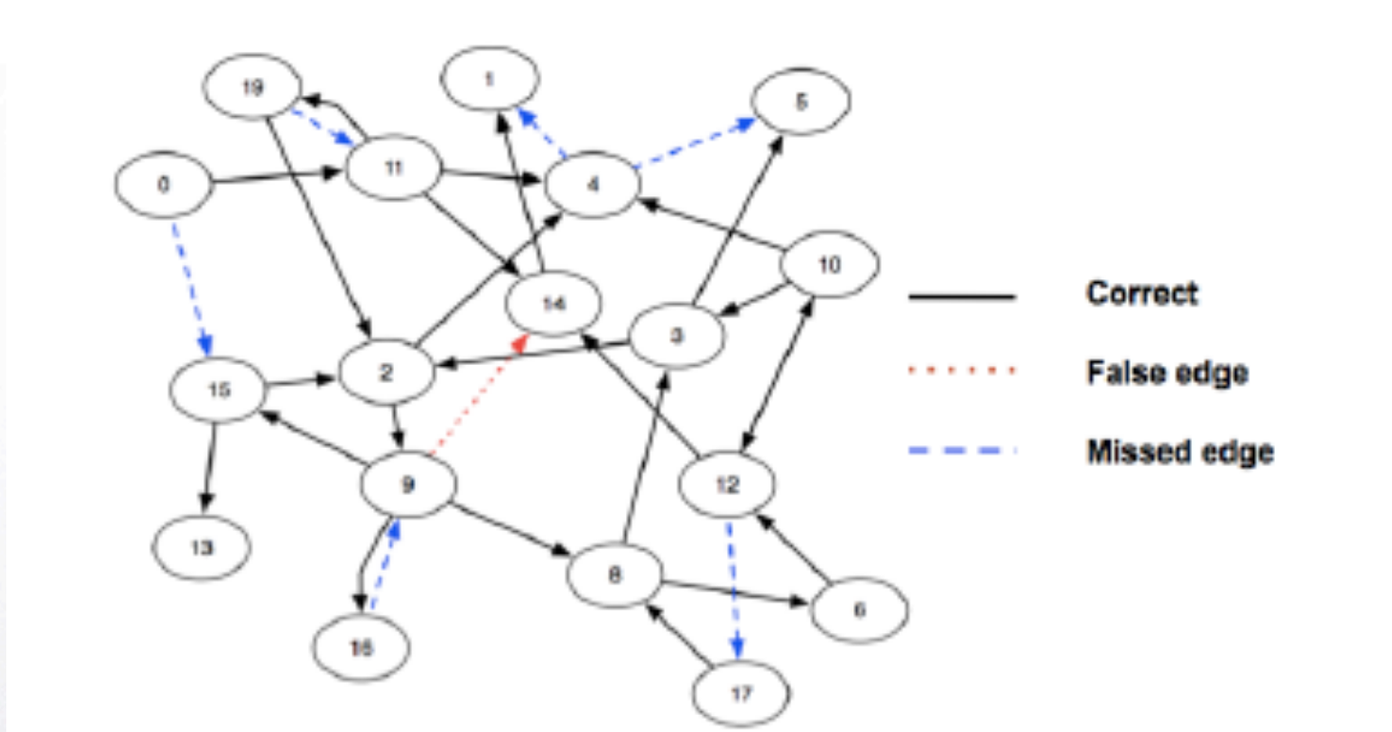
Experiments with synthetic data

Model II

$$\lambda_Y(t|S^t_{\mathcal{N}(Y)}) = \mu + \sum_{X \in \mathcal{N}(Y)} \gamma_X \sum_{t_i \in S^t_X} g(t - t_i)$$

- activity distribution for many users

Recovered network structure for Model II



Working with Twitter dataset

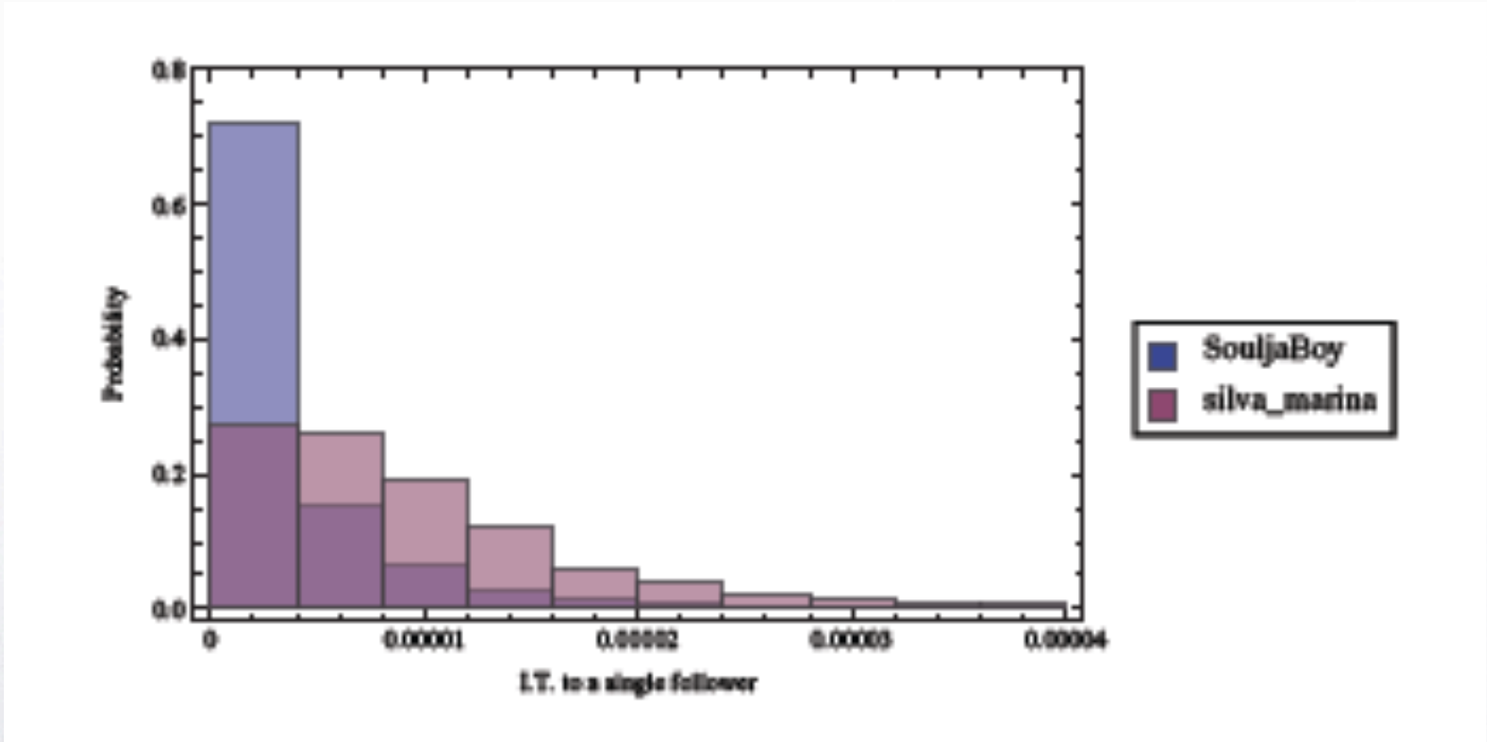
70 000 distinct URLs
3 500 000 tweets
800 000 users
Time period (T) = 3 weeks

Active user tweeted > 10 tweets for three weeks

δ_0 - 1 sec
δ_1 - 10 min
δ_2 - 2 hours
δ_3 - 24 hours

How can we use the information transfer meaningfully?

Probability distribution of outgoing transfer entropy
for SouljaBoy and silva_marina



Conclusions

- Information transfer based approach works without knowing the maps of the relationships $\longrightarrow$ recovering the network structure and finding influentials
- For better estimation many effects impact are taken into account
- Synthetic data based experiments showed rather good results of recovering nertwork structure
- Twitter data analysis proved to be more certain tool for finding the influentials and provided the idea that influence changes through times

Thank you for your attention!