Mining the Information Propagation in a Network

Michele Berlingerio, Michele Coscia, Fosca Giannotti

KDD-Lab, ISTI-CNR Pisa

IMT Lucca Institute for Advanced Studies

Department of Computer Science – University of Pisa

Introduction

- Given a network of users that exchange information,
  - How does the information propagate in a network?
  - Why?
  - How fast?
- Focus on
  - Temporal dimension: topics spread faster than others, distribution of temporal intervals
  - Causes of such spread: characteristics of the users and the topics passed on

Problem Definition

- Dataset D of users U with flow of information as set of timestamped sequences S
  - □ Find frequent patterns of information propagation
  - □ Let the causes of such patterns emerge from the data
- TAS (Temporally Annotated Sequences) mining plus
- Graph Mining

Analysis steps

1. Building a graph G of users U connected by edges representing topics
2. Assigning labels L to U according to their semantical and statistical properties
3. Assigning labels to the edges
4. Extracting flows of information in D
5. Extracting TAS
6. Extracting frequent subgraphs in G

Goal: combining the analysis of the results in 5 and 6

Case study

Data

- Enron emails: after cleaning, 12,000 emails
- 20 newgroup emails: after cleaning, 18,000 email

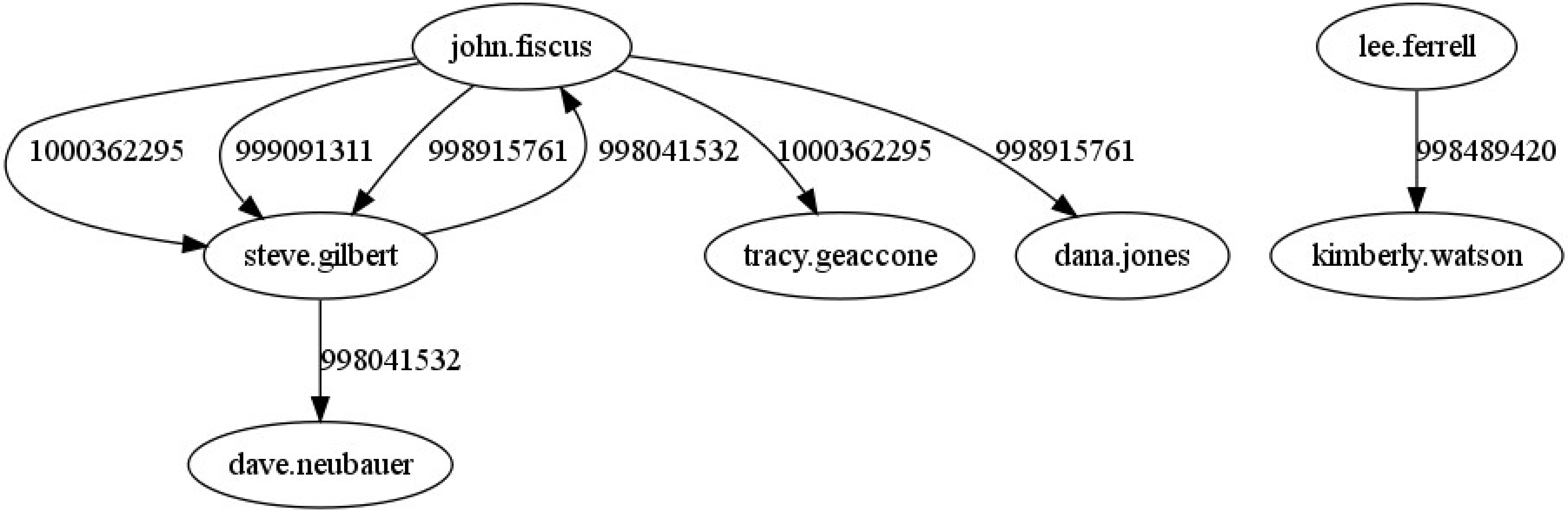
Tools

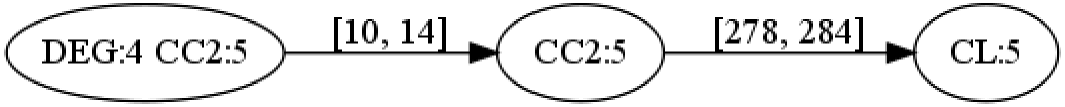
- MiSta software for TAS mining
- Single graph miner

First steps

- Users connected by topics discussed among them
- Users labeled by degree, betweenness, closeness centrality, ..
- Edges labeled by
  - □ words in the topics manually semantically clustered, then label=most frequent cluster
  - □ most frequent word
- For the TAS, emails grouped by subject

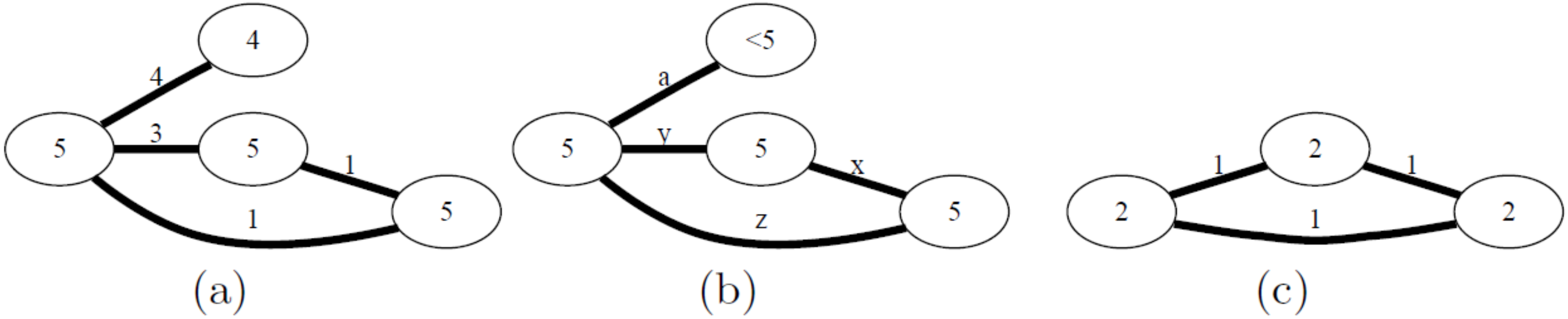
TAS example


john.fiscus
lee.ferrell
1000362295
999091311
998915761
998041532
1000362295
998915761
998489420
steve.gilbert
tracy.geaccone
dana.jones
kimberly.watson
998041532
dave.neubauer

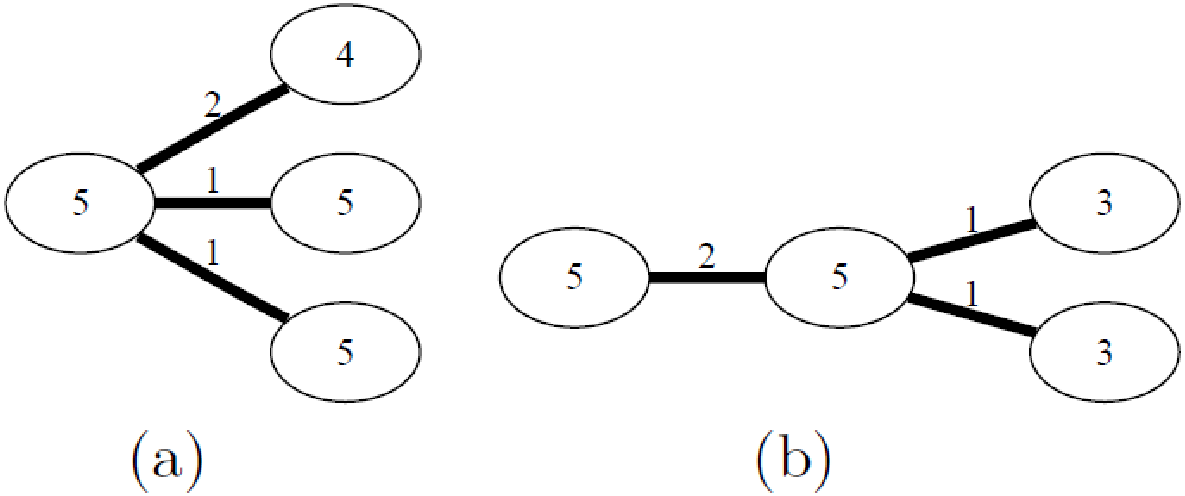

DEG:4 CC2:5
[10, 14]
CC2:5
[278, 284]
CL:5

Patterns found

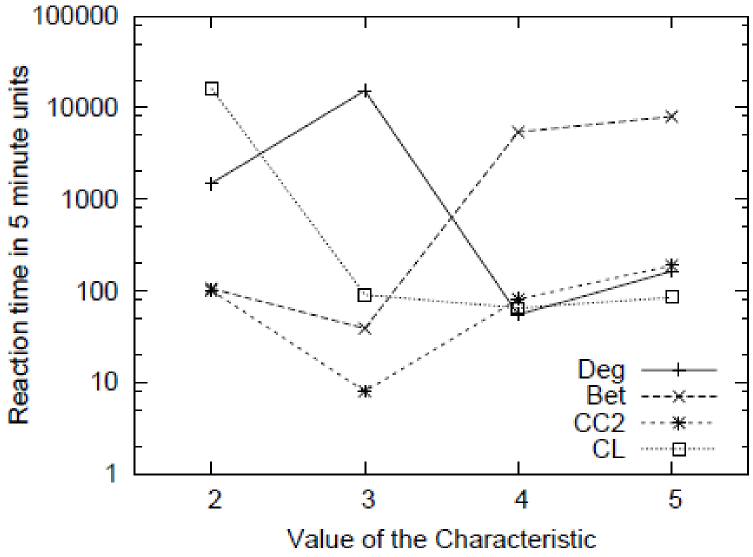
Enron – node labels: CC – edge labels:
most frequent topic (semantically clustered)



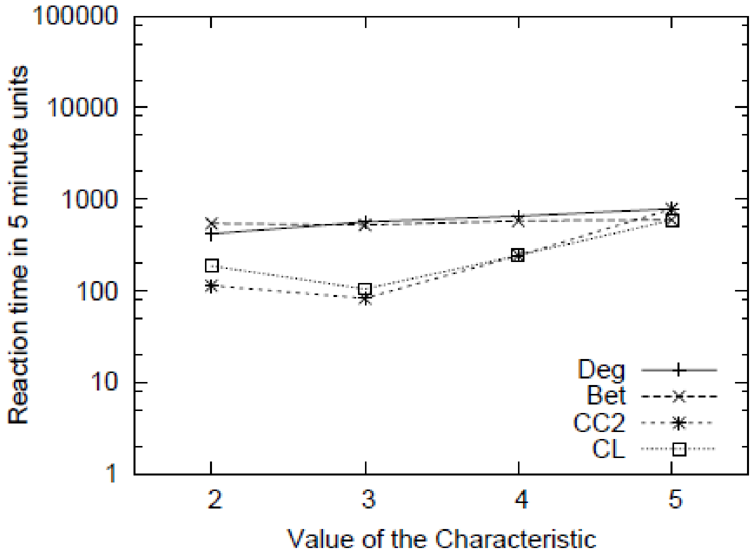
Newsgroup – node labels: CC – edge labels: word frequency



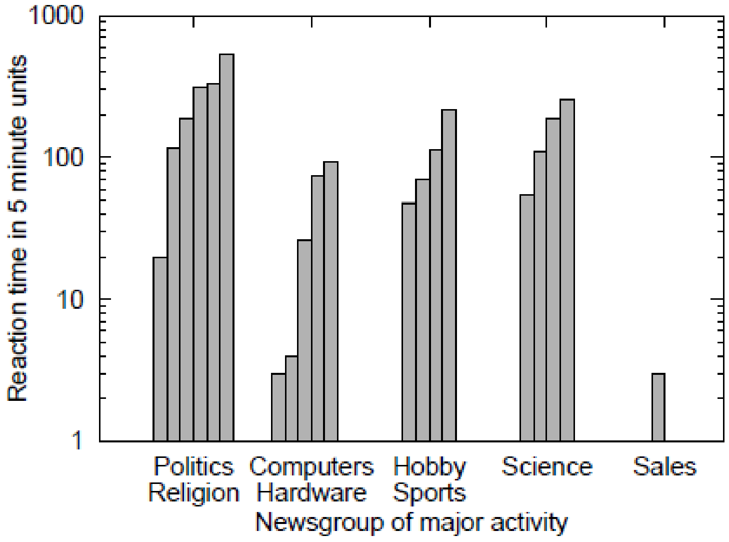
Reaction Times



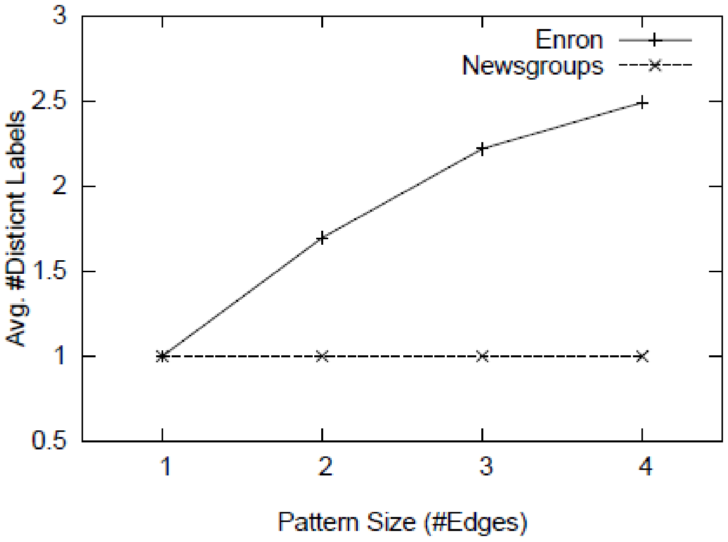
(a) Reaction times - Enron



(b) Reaction Times - Newsgroup



(c) Reaction times - Newsgroup



(d) Graph Patterns Heterogeneity

Conclusions & Future Work

- Extending the case study
- Pushing more semantic on labels
- Comparing different datasets
- Apply the methodology to some real scenario (Viral Marketing, ..)