Generative Models for Rapid Propagation of Information

Kirill Dyagilev (Technion & IBM)

Shie Mannor (Technion)

Elad Yom-Tov (IBM)

Social Networks

The accessibility of large-scale social data lead to an explosion of research in the field of complex networks.

Social data can be used for the following purposes:

- Marketing Campaign management (Hill et.al.)
- Fraud detection (Hill et.al.)
- “Churn” prediction (Nanavati et.al., Richter et.al.)

Influential Subscribers

- One of the central questions - identification of influential subscribers in the network.
  - These subscribers can be used as seeds in marketing campaigns, sources of news items etc.
- Goldenberg et.al. showed a significant role of well-connected individuals in disseminating information and in adoption of innovations.
  - However, he considered a **static** graph of social relations, rather than dynamics of social interaction.

Our contribution

- We investigate the dynamics of information propagation, i.e., the actual sequences of information-passing events.
- We introduce a notion of significance of nodes based on their dynamic behavior.

Rapid Propagation of Information (“Gossip”)

- We focus on **rapid** propagation of information (RPI).

- We look for a sequences of interactions in which once the information is received, it is
  - either transferred to somebody else during a relatively short period of time (say T); or
  - It will not be transferred to anyone.

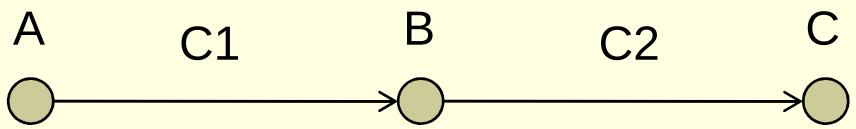
Additional Scenario of Gossip Propagation

Outline

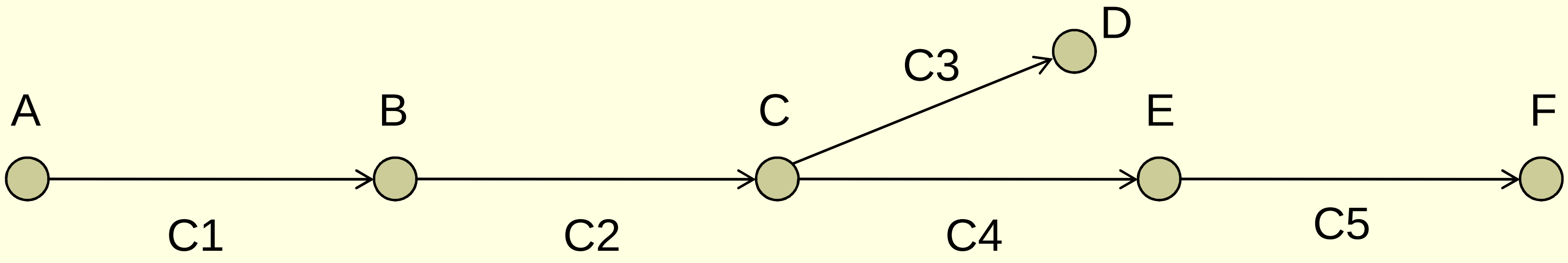
- **Algorithm for identification of event of rapid propagation of information**
- Observations in Real-World data
- Evidence for Information Propagation
- Generative Models of Information Propagation
- Future Work

Rapid Propagation of Information

- Goal: Identify an RPI - sequences of calls involved in rapid propagation of information.
- Calls C1 and C2 are **T-connected** if they share a common subscriber and the time interval between them < T min.

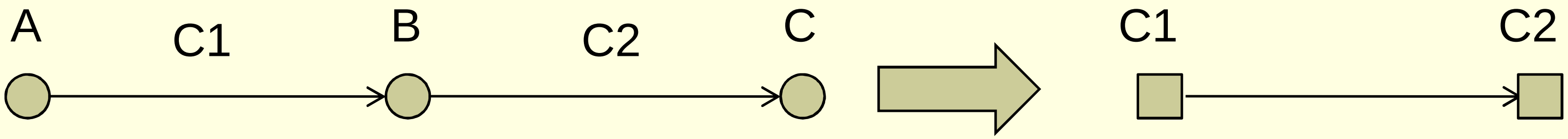


- This observation scales up easily to several calls.



Identification of RPI in Call Data

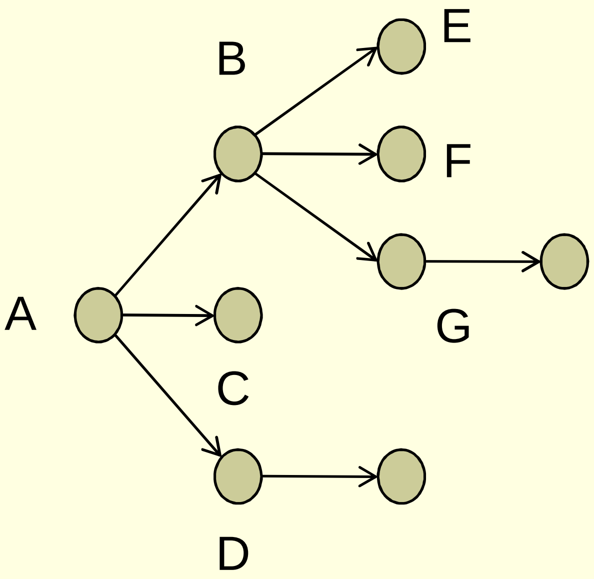
- Build a line graph in which nodes correspond to calls and directed edges connect calls from the same RPI.



- Partition this graph to trees using the DFS algorithm.
- Define large-enough DFS trees (> 4 calls, > 4 subscribers) as RPIs.

Interpretation of GPCs – Information Cascades

- We then translate the set of calls in each RPI to an **information cascade**.
- Namely, we produce a tree that describes paths in which the information propagates from the source subscriber to all the others.



Outline

- Algorithm for identification of event of rapid propagation of information
- **Observations in Real-World data**
- Evidence for Information Propagation
- Generative Models of Information Propagation
- Future Work

Real-world data

- We applied our algorithm to call data records (CDRs) of two large cellular operators from different parts of the world:

Operator 1:

- 50 million calls over 24 days,
- total 5.4 million of distinct subscribers, out which approximately 2 million belonged to the analyzed operator.

Operator 2:

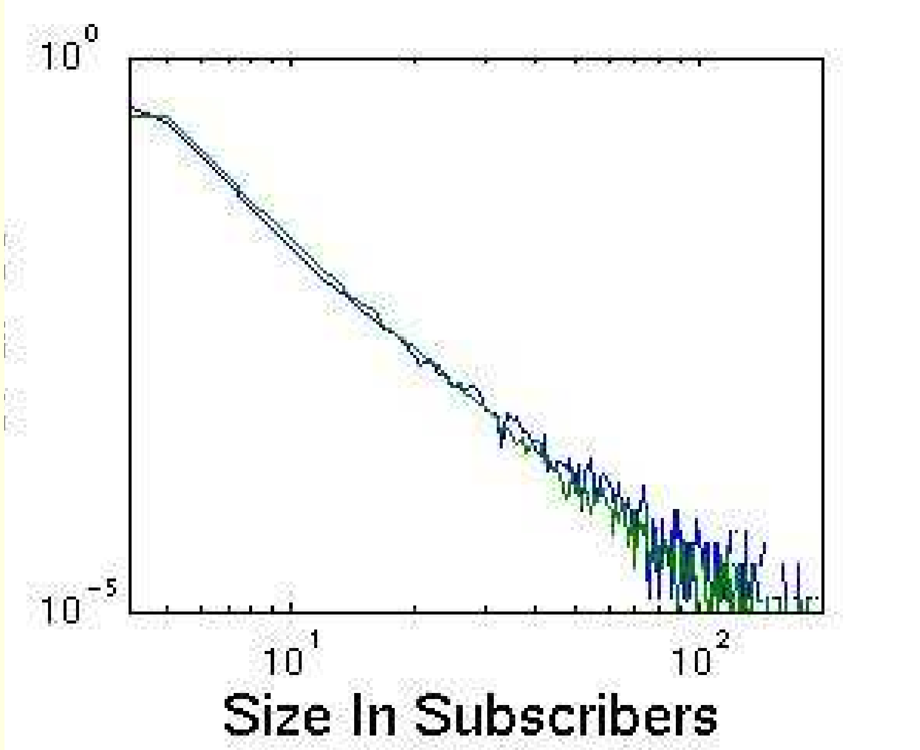
- Twice as many calls in the same period of 24 days.
- Similar number of subscribers.

Real-world data (cont.)

- Description of each call contains:
  - Obfuscated identity of subscribers involved.
  - Beginning time of the call and its duration.

Structural Properties of RPIs

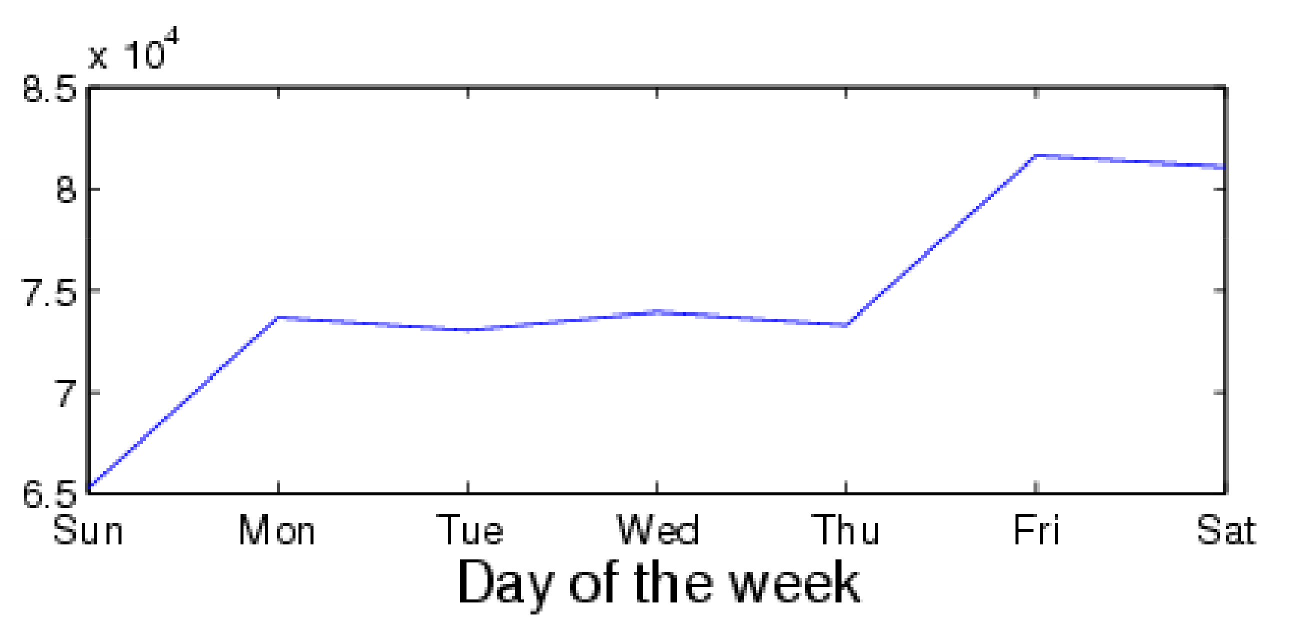
- Size distribution of RPIs (T=20min):



- Size distribution is almost identical for both data sets.

Structural Properties of RPIs

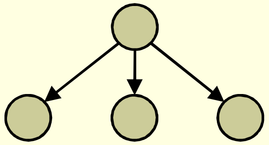
- Average number of RPIs by weekdays (T=20min):



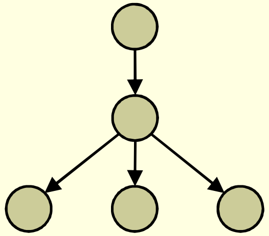
Properties of Information Cascades

We used clustering to isolate typical topologies of information cascade.

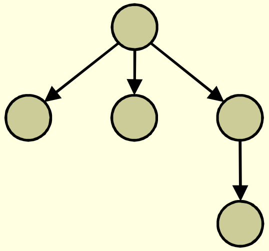
1. Pure star.



2. Initialization call + pure star.



3. Pure star + single additional node.

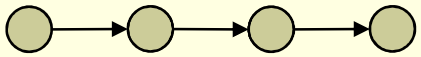


These topologies cover over 60% of all RPIs.

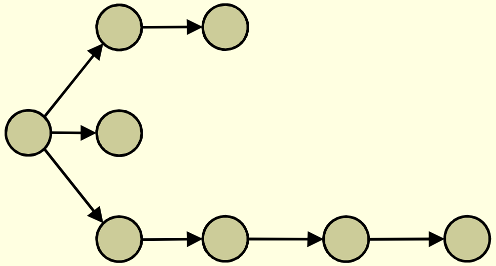
They all have one dominant node – **dissemination-leader.**

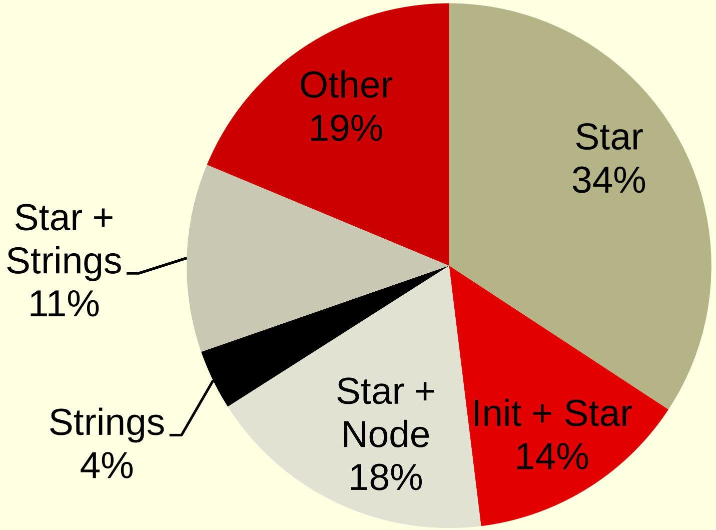
Properties of Information Cascades (cont.)

4. Strings.



5. Star + Strings.





6. The rest of the trees.

Dissemination-Leaders Vs. Hubs

- We compared the set of hubs (subscribers with top 5% of number of friends) and the set of dissemination-leaders.

- These sets overlap, but differ in a significant way:
  - 41% of hubs are also dissemination-leaders.
  - 64% of dissemination-leaders are hubs.

Outline

- Algorithm for identification of event of rapid propagation of information
- Observations in Real-World data
- **Evidence for Information Propagation**
- Generative Models of Information Propagation
- Future Work

Do RPIs really propagate information?

- Downside: without knowing the content of calls, it is impossible to verify that RPIs disseminate information.

- Upside:
  - RPI cover several intuitive scenarios of information propagation.
  - Basic properties of RPIs make sense.
  - We can provide certain circumstantial evidence for the hypothesis.

Geographic Evidence for Information Propagation

- The following experiment shows that some RPIs propagate geospatial information.
- We can estimate the location of a subscriber using the number of the antenna (cell) his phone uses during the current call.
- Consider cells visited in a single day by a pair of socially connected subscribers: A and B.

A		B	A
A&B			B
		A&B	B

Geographic Evidence for Information Propagation

- Consider 85,000 pairs of socially-connected subscribers

- Count the number of “shared” cells
  - On a day in which they appeared in the same RPI.
  - On a day their communication did not appear in a RPI.

- The number of “shared” cells increases on the day these subscribers participate in the same RPI.

Outline

- Algorithm for identification of event of rapid propagation of information
- Observations in Real-World data
- Evidence for Information Propagation
- **Generative Models of Information Propagation**
- Future Work

Propagation Models

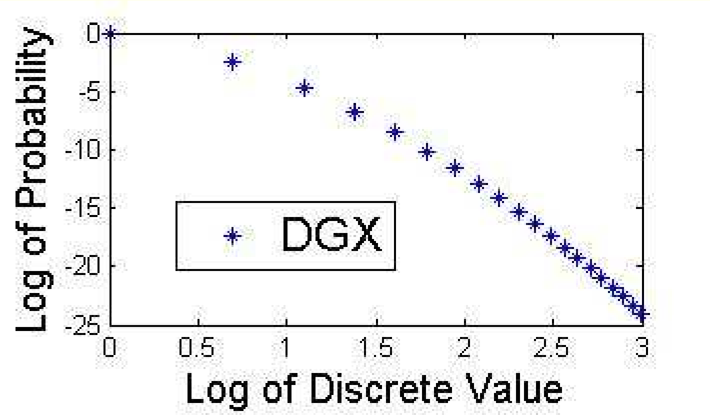
- Day Generating Model:
  - Describes the emergence of sequences of calls that produce RPIs with the given size distribution.
- Information Cascade Model:
  - Generates Information Cascades of different topologies.
  - Fits the given fraction of RPIs of each topology and given size distribution.

Day Generating Model - Assumptions

- This model relies on the following assumptions:
  - Two kinds of subscribers: regular and dissemination-leaders.
  - Fraction of dissemination-leaders is relatively small => dissemination-leaders call only regular subscribers.
- The model generates calls made by a dissemination-leader during a single day.
- Resulting topology is simplistic, but covers over 50% of RPIs in data.

Day Generating Model – Some Details

- **Number of calls** is Discrete Gaussian eXponential (DGX)



- **Beginning time of the first call** is uniform over the day.
- **Time interval between consecutive calls** depends on the total number of calls and is DGX.
- **Callees** are chosen uniformly from the set of regular subscribers.

The fit of the Day Generation Model to data

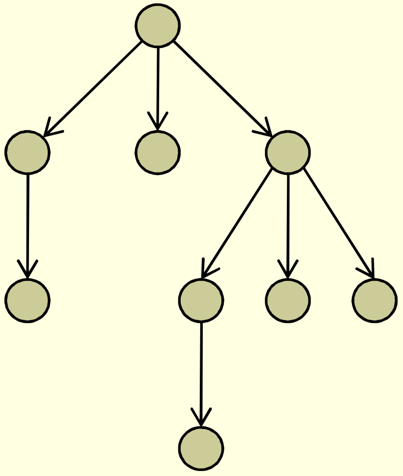
- This model explains well the size distribution of RPIs (R-squared = 0.88).

- The model admits combinatorial analysis => size distribution can be predicted theoretically.

Information Cascade Model

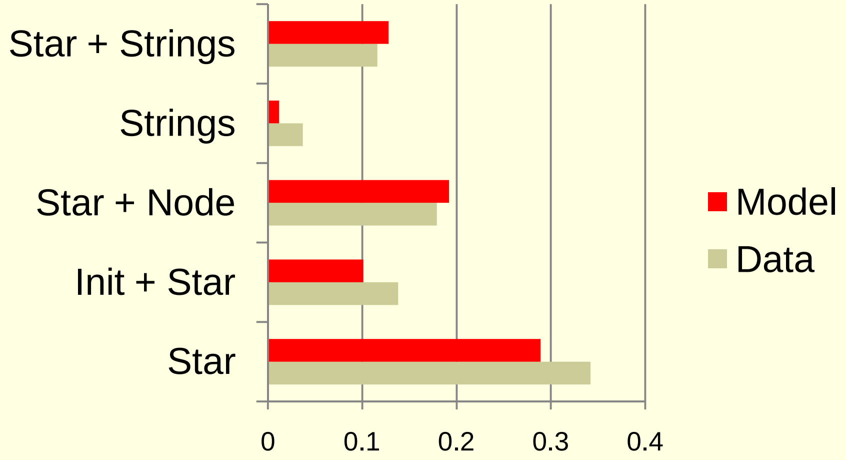
- We use branching process to model the information cascade, namely, the corresponding tree is built in a layer-by-layer fashion.

- Degree distributions are modeled by Discrete Gaussian eXponential (DGX) and depend on the following properties:
  - depth of the current node
  - degree of the root



The fit of the Information Cascade Model to data (cont.)

- The information cascade model predicts the fraction RPIs belonging to each topology.
  - Both using theoretical results and simulation



- This model explains well the size distributions of RPIs of different distributions (R-squared > 0.95).

Outline

- Algorithm for identification of event of rapid propagation of information
- Observations in Real-World data
- Evidence for Information Propagation
- Generative Models of Information Propagation
- **Future Work**

Future Work

- More circumstantial evidence for information propagation.
- Model unification: generation of sequences of calls that disseminate information and the topology of the information cascades.
- Inter-day behavior of dissemination-leaders.
- Apply our approach to other media, e.g., twitter.