Spinner detection

Industry: Real Impact Analytics

Industry Representative: Sebastien Leempoel

Team: Dr. Turgay Celik, Dr. Kamil Kulesza, Michael Mitchley

Mathematics in Industry Study Group 2014

Content

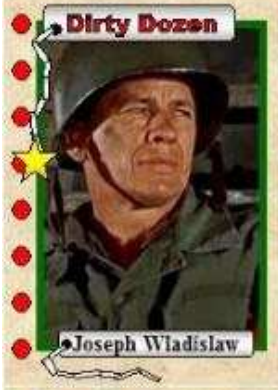
- Intro
- Problem description and motivation
- Information collected
- Data: synthetic and real
- Graph based approach (social networks)
- Bayesian approach
- Summary and further work

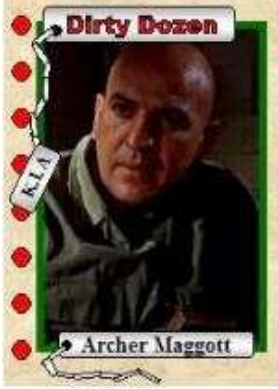
- **Intro**
- Problem description and motivation
- Information collected
- Data: synthetic and real
- Graph based approach (social networks)
- Bayesian approach
- Summary and further work

The Dirty Dozen file

classified

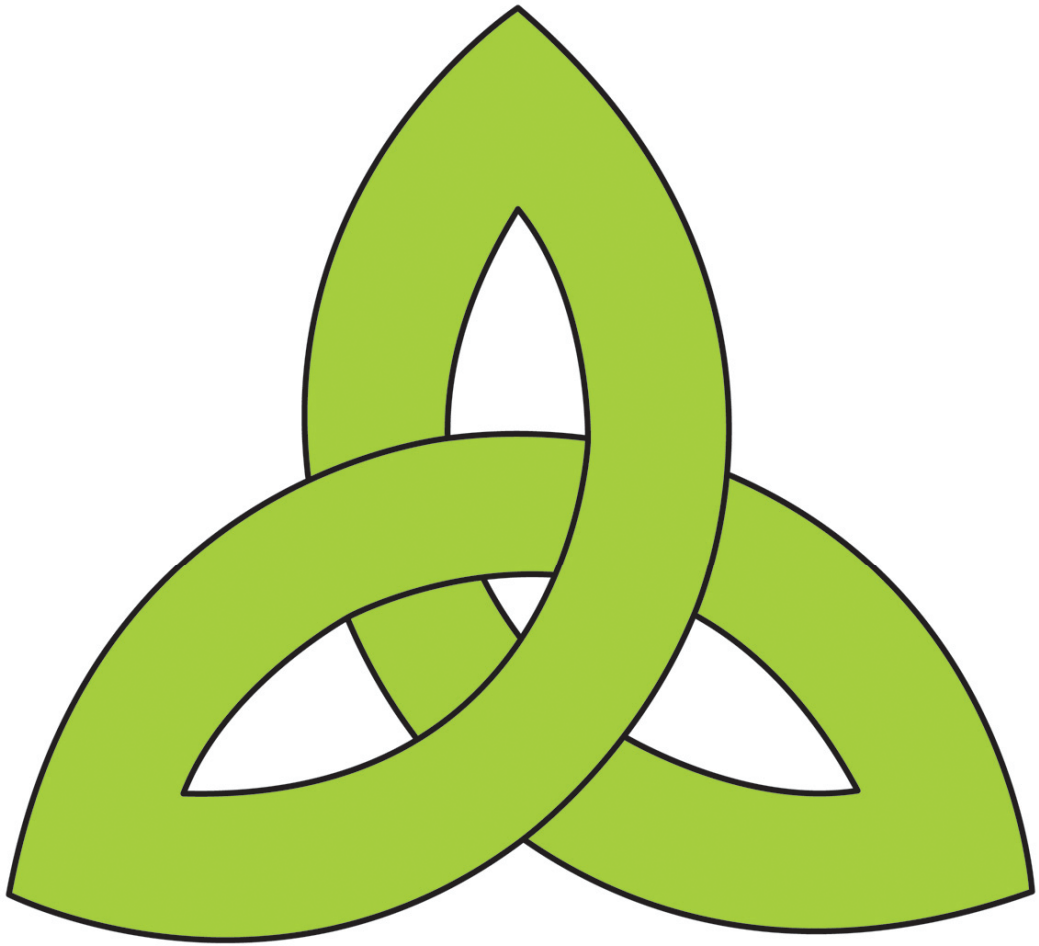
The War Collection





SEVEN SAMURAI


TEENAGE MUTANT NINJA
TURTLES



TRINITY

- Intro
- **Problem description and motivation**
- Information collected
- Data: synthetic and real
- Graph based approach (social networks)
- Bayesian approach
- Summary and further work

Mobile networks business model

Spinners phenomenon and its impact on the business model (FAR vs FRR)

Find spinners (to keep them longer in the network)!

Tasks as given at the beginning (wish list):

1. **Formalize** the problem in mathematical terms. In particular, the team will need to decide how to measure **similarity of social networks** and mobility between two users;
2. Suggest and implement an algorithm to detect spinners on a small **synthetic set** of Call Data Records (CDRs), which will be provided by Real Impact;
3. **Test** their method on a large synthetic set of CDRs and adjust their method to computational constraints;
4. (*optional*) if their method performs correctly (both in speed and results) on the large synthetic CDRs, their method will be tested by Real Impact on real-world data, to observe the behaviour of the method in real-market situations. The performance of the method will be shared with the team (for con fidentiality reasons, direct access to real-world data won't be possible).

- Intro
- Problem description and motivation
- **Information collected**
- Data: synthetic and real
- Graph based approach (social networks)
- Bayesian approach
- Summary and further work

Selected information acquired

- Contract and pre-paid subscribers
- Subscribers' identification
- Multi-SIM-ers
- Churners and spinners
- Promotions made by SA mobile provider
- Promotions as a mean to acquire additional information

- Intro
- Problem description and motivation
- Information collected
- **Data: synthetic and real**
- Graph based approach (social networks)
- Bayesian approach
- Summary and further work

Synthetic data

day_index	A	B	cell_A	
1	18	36	48	
1	16	62	13	
1	98	7	39	

Real data

DATE_TIME: date and time of event, represented as a string: information on date, hour, minute and second of call intitation (format may vary)

DIRECTION: Incoming or outgoing. If incoming, the meta-data are for the called party. If outgoing, the meta-data are for the calling party.

A: number of calling party (international prefix can help for country detection and possibly local prefix to detect the operator (some countries)

B: number of called party (idem)

TYPE: Voice, SMS, Data (sometimes MMS)

VOLUME: amount of units consumed (seconds for calls, Kb for Data, not In use for SMS)

VALUE: value charged to the customer (calling party if outgoing, called party if incoming)

CELL ID: Cell ID of where the customer (calling party if incoming, called party if outgoing) is located when the call was initiated

CUSTOMER_TYPE: type of customer (calling party if incoming, called party if outgoing): usually prepaid or postpaid

Synthetic data

day_index	A	B	cell_A	
1	18	36	48	
1	16	62	13	
1	98	7	39	

Real data

DATE_TIME: date and time of event, represented as a string: information on date, hour, minute and second of call intitation (format may vary)

DIRECTION: Incoming or outgoing. If incoming, the meta-data are for the called party. If outgoing, the meta-data are for the calling party.

A: number of calling party (international prefix can help for country detection and possibly local prefix to detect the operator (some countries)

B: number of called party (idem)

TYPE: Voice, SMS, Data (sometimes MMS)

VOLUME: amount of units consumed (seconds for calls, Kb for Data, not In use for SMS)

VALUE: value charged to the customer (calling party if outgoing, called party if incoming)

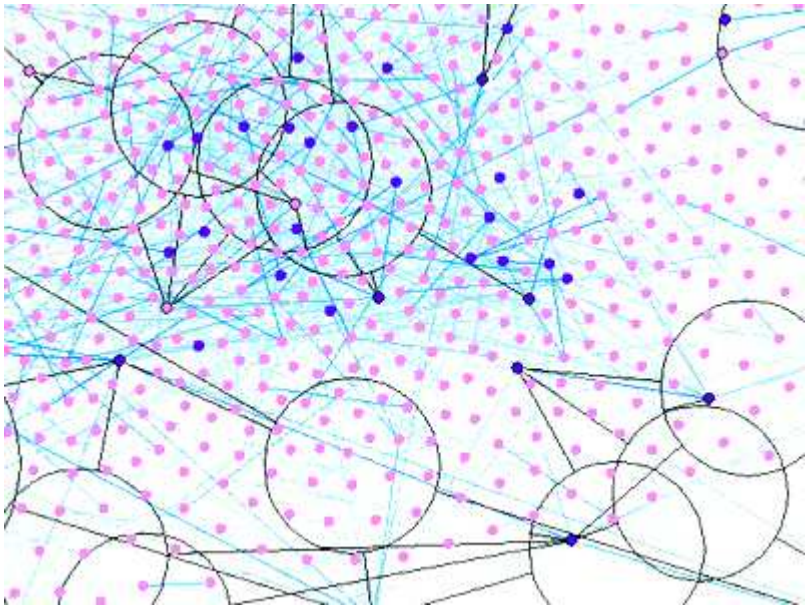
CELL ID: Cell ID of where the customer (calling party if incoming, called party if outgoing) is located when the call was initiated

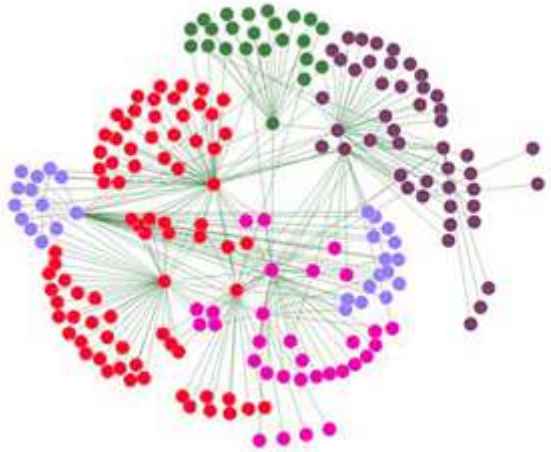
CUSTOMER_TYPE: type of customer (calling party if incoming, called party if outgoing): usually prepaid or postpaid

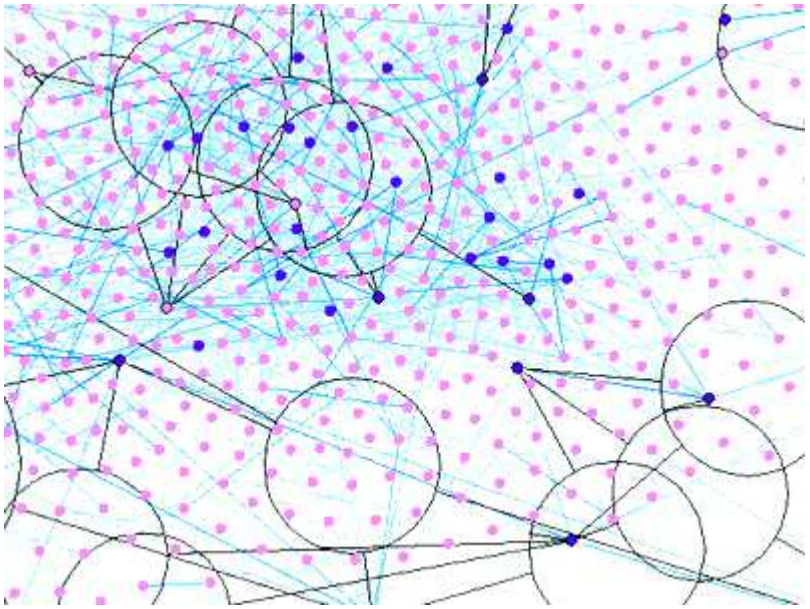
some remarks on: - synthetic data validation

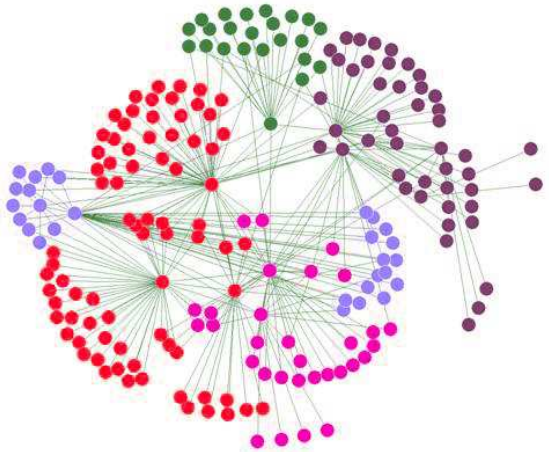
- spatial and temporal information

- Intro
- Problem description and motivation
- Information collected
- Data: synthetic and real
- **Graph based approach (social networks)**
- Bayesian approach
- Summary and further work









Suggested use of graph theory and social networks – some remarks

- Intro
- Problem description and motivation
- Information collected
- Data: synthetic and real
- Graph based approach (social networks)
- **Bayesian approach**
- Summary and further work

Notation

x	An arbitrary SIM card currently on the network.
x_k	An arbitrary SIM card historically on the network.
$E(x)$	The evidence or data associated with x.
$C(x_k)$	The SIM card x_k is a *Churner*.
$S(x)$	The SIM card x is a *Spinner*.

Problem formulation

The question: what is the probability that x matches x_k given the evidence of both, and the knowledge that x_k is a churner?

$$P(x = x_k | C(x_k) \wedge E(x) \wedge E(x_k)).$$

Through Bayes' Theorem, this is equivalent to

$$P(x = x_k | C(x_k) \wedge E(x) \wedge E(x_k)) = \frac{P(E(x) \wedge E(x_k) | x = x_k \wedge C(x_k)) P(x = x_k | C(x_k))}{P(E(x) \wedge E(x_k))}.$$

Now,

$$\begin{aligned} P(E(x) \wedge E(x_k)) &= \sum_{i=0}^{n} P(E(x) \wedge E(x_i) | x = x_i \wedge C(x_i)) \\ &= \sum_{i=0}^{n} P(E(x) | x = x_i \wedge C(x_i)) P(E(x_i) | x = x_i \wedge C(x_i)) \end{aligned}$$

Looking at one term

$$P(E(x)|x = x_i \wedge C(x_i))$$

is the probability of seeing the evidence associated with x, given that x matches a known churner x_i. This can be computed from the data , since both $E(x_i)$ and $E(x)$ are collections of samples drawn from some hidden, underlying distribution.

$$P(E(x_i)|x = x_i \wedge C(x_i)) = P(E(x_i)|C(x_i))$$

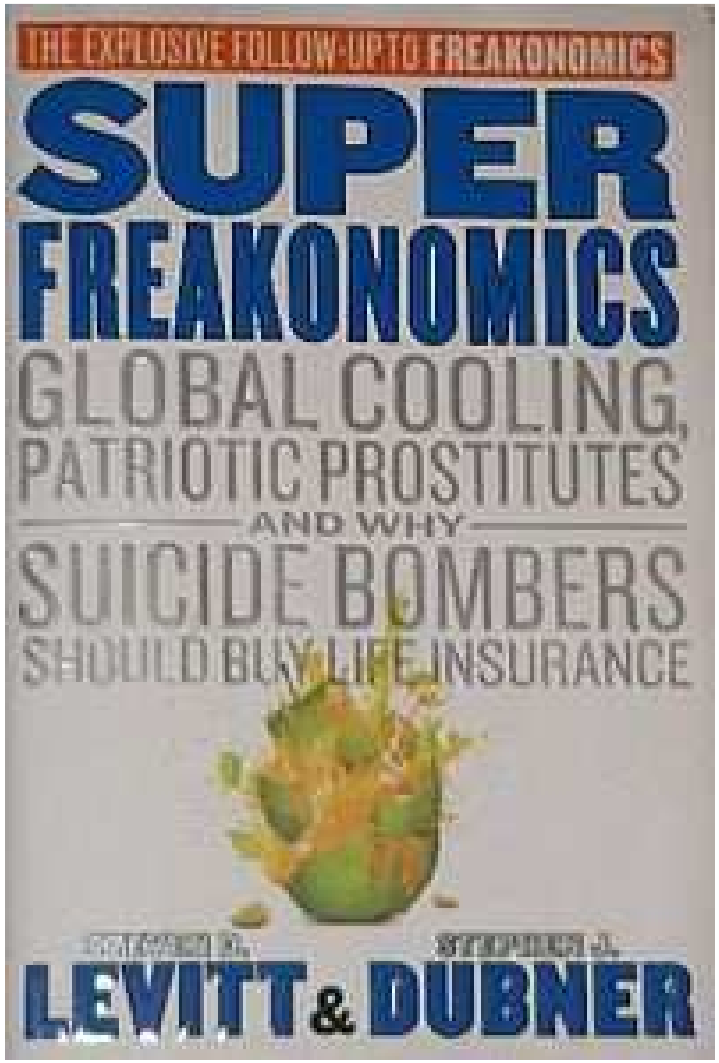
is then a confidence measure of our historic data, since this is the probability of seeing evidence $E(x_i)$ given that x_i is a churner.

$$P(x = x_k|C(x_k))$$

is simply the likelihood that x matches x_k in the absense of evidence, and for n churners ought to be $\frac{1}{n}$.

some remarks on evidence/signature creation and refinement

Bayesian approach in similar applications (friendly example)



- Intro
- Problem description and motivation
- Information collected
- Data: synthetic and real
- Graph based approach (social networks)
- Bayesian approach
- **Summary and further work**

Summary

keywords from the wish list:

formalize, similarity of social networks, synthetic data set, test

graph and Bayesian based approaches

Further work

synthetic data sets

studying spinners

evidence and signatures

other approaches (eg. Remember FAA vs FRR)?