Predictive Analytics with Social Network Analysis (SNA)

Frank Vanden Berghen

Business insight

http://www.business-insight.com

Definitions of the main terms in the SNA datamart (1/2)

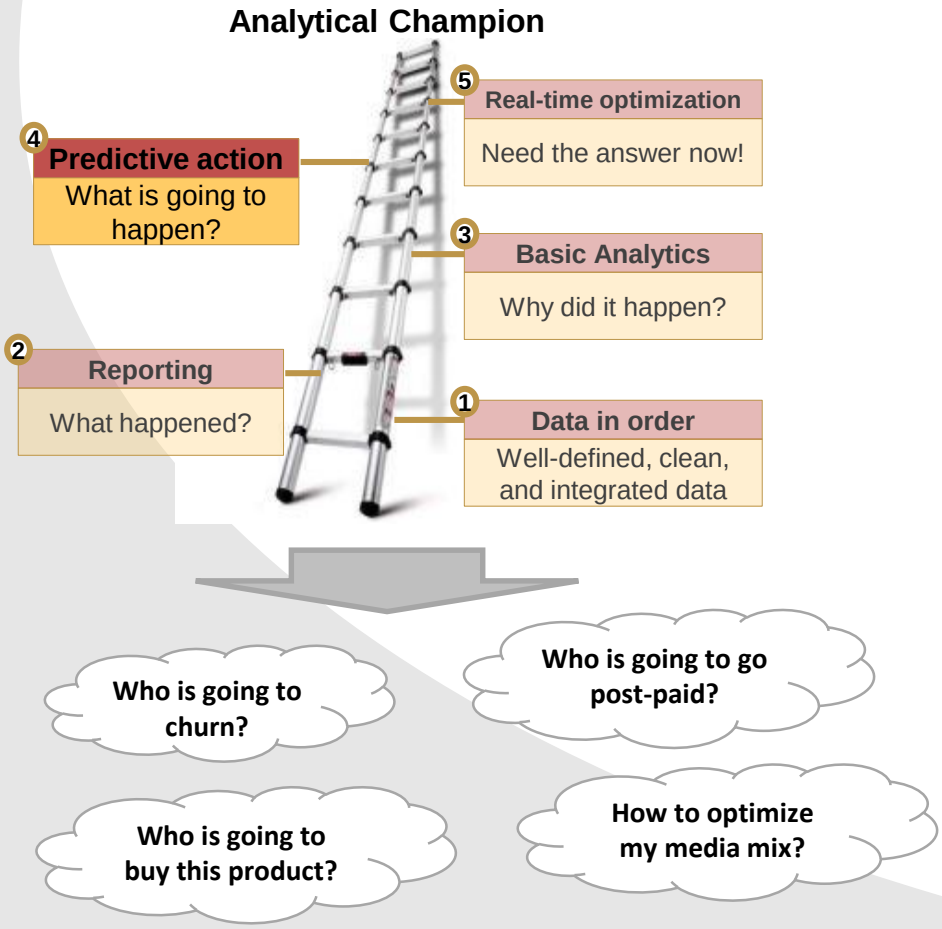
Indicator	Definition
Churn	• Dormant 15 (Subs who has had no outgoing activity in the last 15 days – SMS, Voice, data)
Churn propensity	• The likelihood of current subscribers to become Dormant 15 in the next 15days
SNA Weight	• Weight is proportional to number of SMS (in/out) and total MOU of Voice calls (in/out) during the home hours
Community	• Group subscribers who are more connected between themselves than the outside world • Subs are assigned into different communities such that it maximises the modularity value of the graph. The modularity value is the fraction of the edges that fall within the community minus the expected value of the same quantity if arcs fall at random without regards for the community structure (see appendix for more details)
Community of communities	• Group of subscribers who are more connected between themselves than the outside world (based on the concept of modularity - see appendix for more details)
Social Leader	We use two methods: • Centrality: Node with the highest centraility score where the centrality of a node in the network is computed as the sum of all paths going through that node, weighted by their length • Social degree: Node with the highest social degree where the social degree of a Subs is defined as the number of triangles made by them in the graph (see appendix for more details)
Distance to SL	• -1 if not connected to a SL, 0 for a social leader then +1 for each hop that separates the Subs to the SL
Significant Other	• Top people with the highest SNA weight and the same most used cell during home hours
Close Friend	• Top 5 people with the highest SNA weight
Social Value	• Revenue of the Subscriber plus 12% of the SO revenues and 12% of the close Friends
Revenue at risk	• Revenue of the Subs times the Churn propensity
Total value at risk	• Social Value of the Subs times the Churn propensity

Definitions of the main terms in the SNA datamart (2/2)

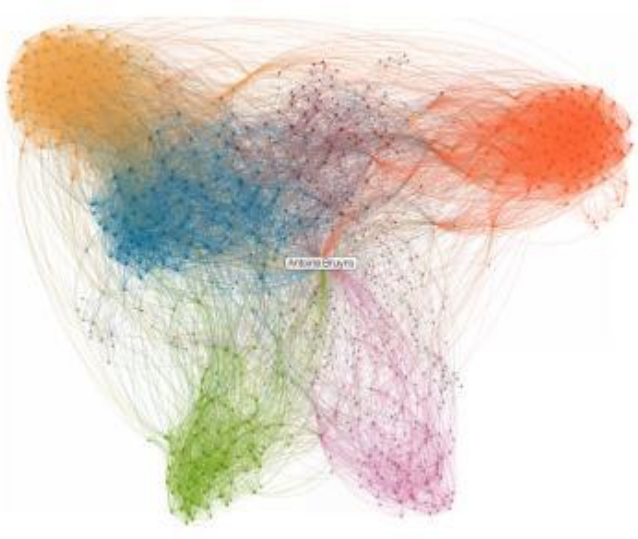
Indicator	Definition
Interval	• Number of days between the observation date and the last outgoing activity (SMS sent, Voice made, data package sent) (no activity incoming vs. outgoing?)
Busy Hour/ Home Hour	• Busy hours are from 6h to 18h
Most used cell	• Cell with the highest number of activities (number of SMS sent/received & number of Voice call sent/received)
Geographic distance to x	• Distance based on the GPS coordinates between two cells
Distance travelled	• Length of the curve linking all the cells used by the Subs over a given period
OnNet	• All the calls between a Tigo Subs and a Tigo Subs (both local and international)
OffNet	• All the calls between a Tigo Subs and a Subs from another local carrier(excl. the international calls)
International	• All the calls between a Tigo Subs and a Subs from another international carrier

Predictive Analytics with Social Network Analysis allows you to answer some of a telco's core business questions with more accuracy

Predictive Analytics helps answer core business questions

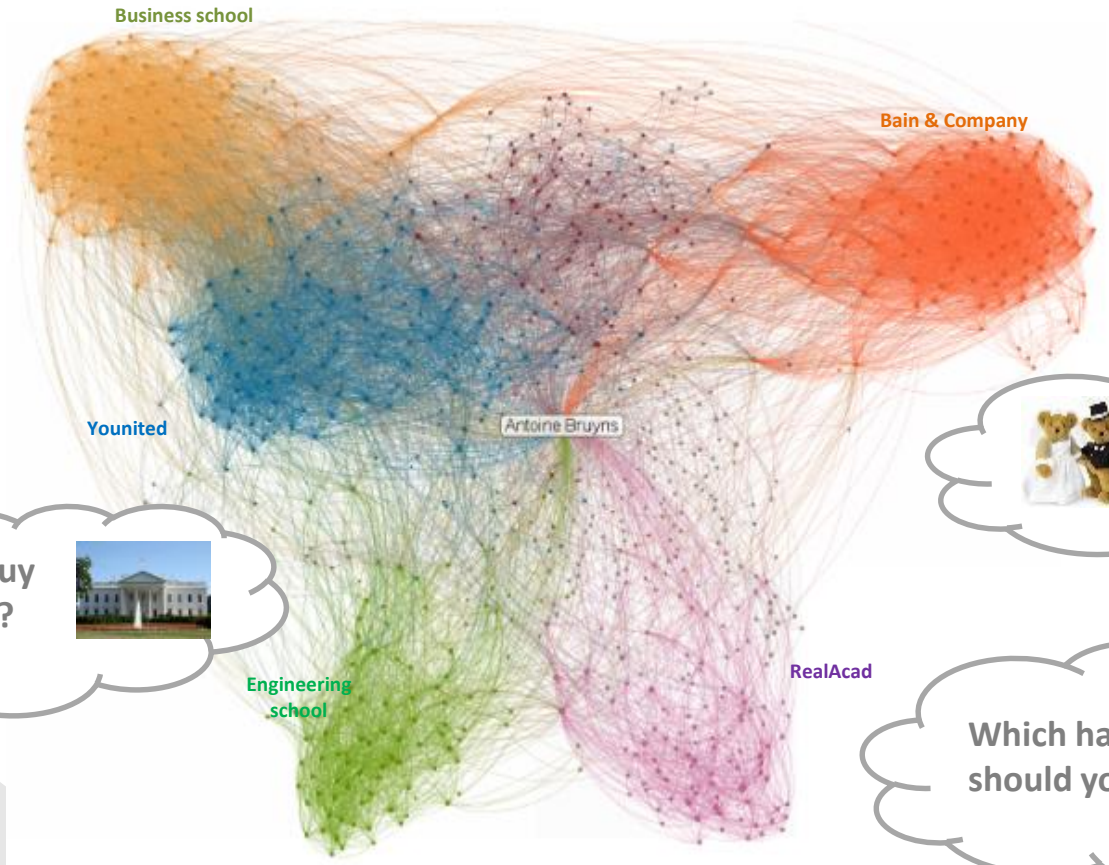


Social Network Analysis reveals new layers of information in your data



- Analyze the interdependency between your Subscribers
- Add extra-information to your predictive models
- More info makes your model more accurate

Social Network Analysis matters because Humans are social creatures; they make decisions based on the impact of their communities



Where to buy your house?

Who should you marry?

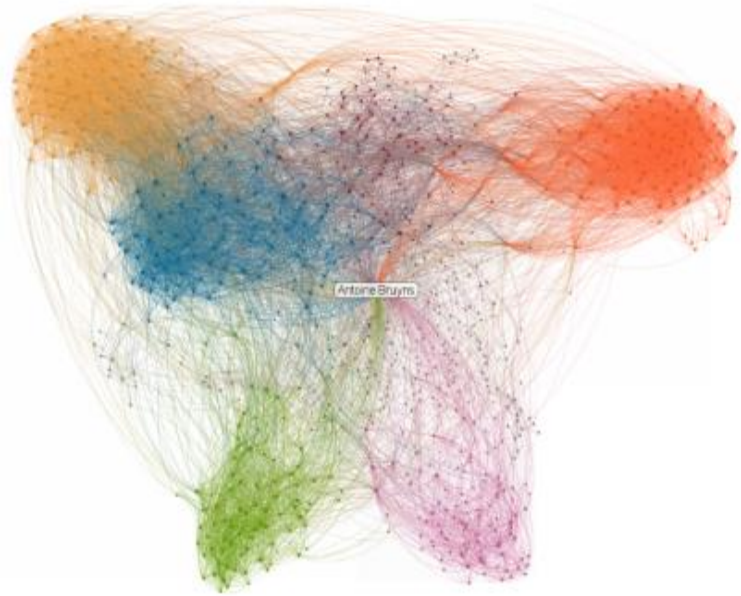
Which handset should you buy?

Core assumption: *"Our group of friends influences us, our choices, our preferences & our decisions"*.

Note: you can build the same map with inmaps ☺

Social Network Analysis is the mathematical analysis of human relationships to understand how we influence each other

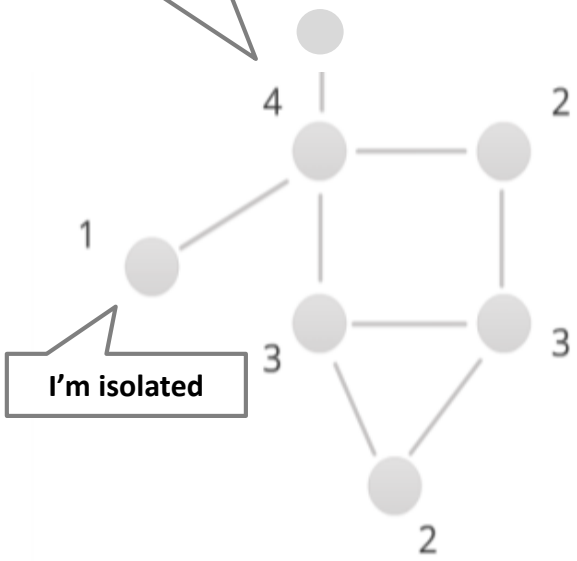
Who influences me?



Who are the connectors?

Who has no influence?

I'm the leader. People listen to me!



I'm isolated

How you build communities & which information you extract depends on your methodology and parameters

Step 1: Build communities
define nodes, links, link weights, etc

Step 2: Derive insights
calculation of relevant information about nodes in the network

- **Nodes** are the people
- **Links** show relationships between the nodes
- **Link weight** is the strength of the "social bond" among two people

- **Clustering coefficient**
  measures the local density around a node
- **Centrality measures**
  quantify the "importance" of a particular node within a network
- **Communities**
  Community detection algorithm builds the best partition of the network in communities based on the concept of **modularity Q**

Community detection methods take into account the global structure of the network to deduce which communities make (most) sense

What is the best partition of this network?

$e_{ij} =$ fraction of all the arcs in the network between community i and j

$e_{ii} =$ fraction of all the arcs in the network inside community i

$a_i = \sum_j e_{ij}$ =fraction of arcs with one extremety in community i

$a_i^2 =$ fraction of arcs with the 2 extremeties in community i

$m =$ number of arcs inside the complete graph

$mQ =$ (number of links in communities) - (number of expected links in communities)

$$mQ = m\left(\sum_{i=1}^{nCommunity} e_{ii}\right) - m\left(\sum_{i=1}^{nCommunity} a_i^2\right)$$

$$Q = \sum_{i=1}^{nCommunity} (e_{ii} - a_i^2)$$

Community detection methods take into account the global structure of the network to deduce which communities make (most) sense

What is the best partition of this network?

Q = (number of links in communities) – (number of expected links in communities)

Gamma=.3

Expected= gamma * max number of arc in comm.

21 3

Q= (13+3)-0.3 x (21+3)=8.8

Q= (13+3)-0.3 x (21+3)=8.8

Q= (3+4+3)-0.3 x (3+6+3)=6.4

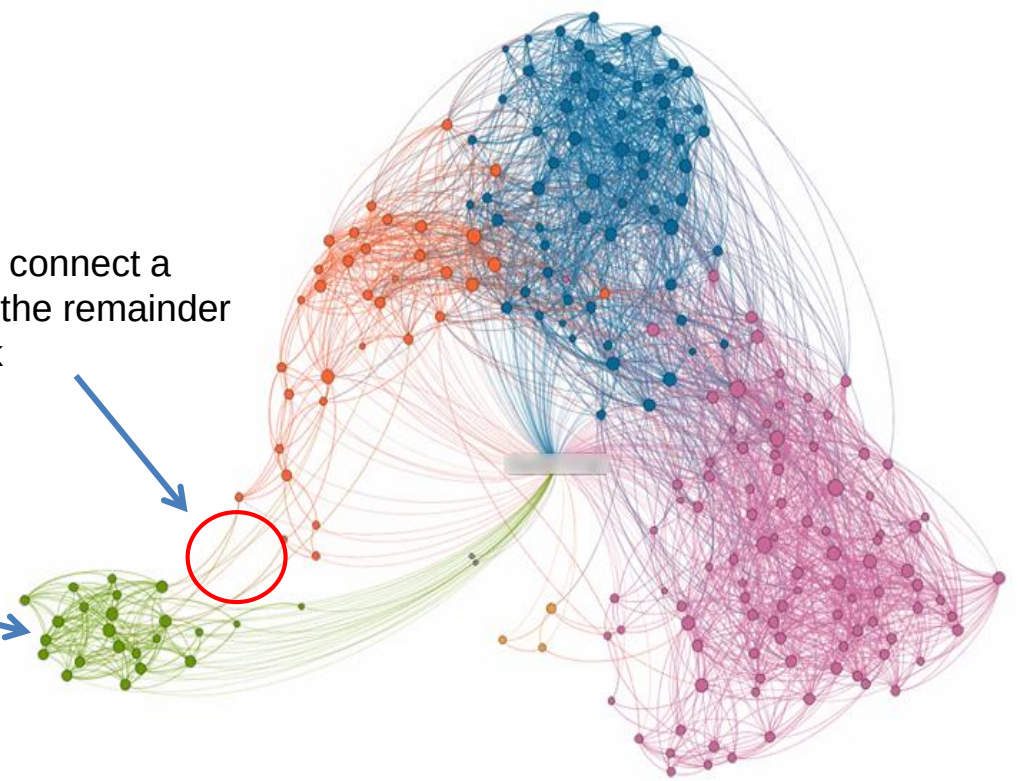
Q= (10+10)-0.3 x (10+10)=14

Communities discovery (1/5)

The detection of communities in a network consists in dividing the network into groups of tightly connected nodes

Only few links connect a community to the remainder of the network

There are many links inside the community (in comparison to a random network)



How do we detect such a partition? Classical clustering methods (Kernighan-Lin, K-means, …) **FAIL** at this

Communities discovery (2/5)
The optimization of modularity provides good communities in a short time (given the right optimization method)

Modularity is defined as the difference between the number of links observed in the community and its expected number in a random model

$$Q = \frac{1}{m} \sum_{i,j \in N} (A_{ij} - \frac{{k_i}^{out} {k_j}^{in}}{m}) \partial(c_i, c_j)$$

Where

- A_{ij} is 1 if the link (i,j) exists, and 0 otherwise
- ${k_i}^{out}$ and ${k_j}^{in}$ are respectively the outdegree of node i and the indegree of node j
- m is the number of links in the network
- $c_i$and c_j are respectively the community index of i and j
- $\partial(\cdot,\cdot)$ is the classical Dirac function

Communities discovery (3/5)
A multi-level optimization method relies on the local properties of modularity to compute the optimal communities

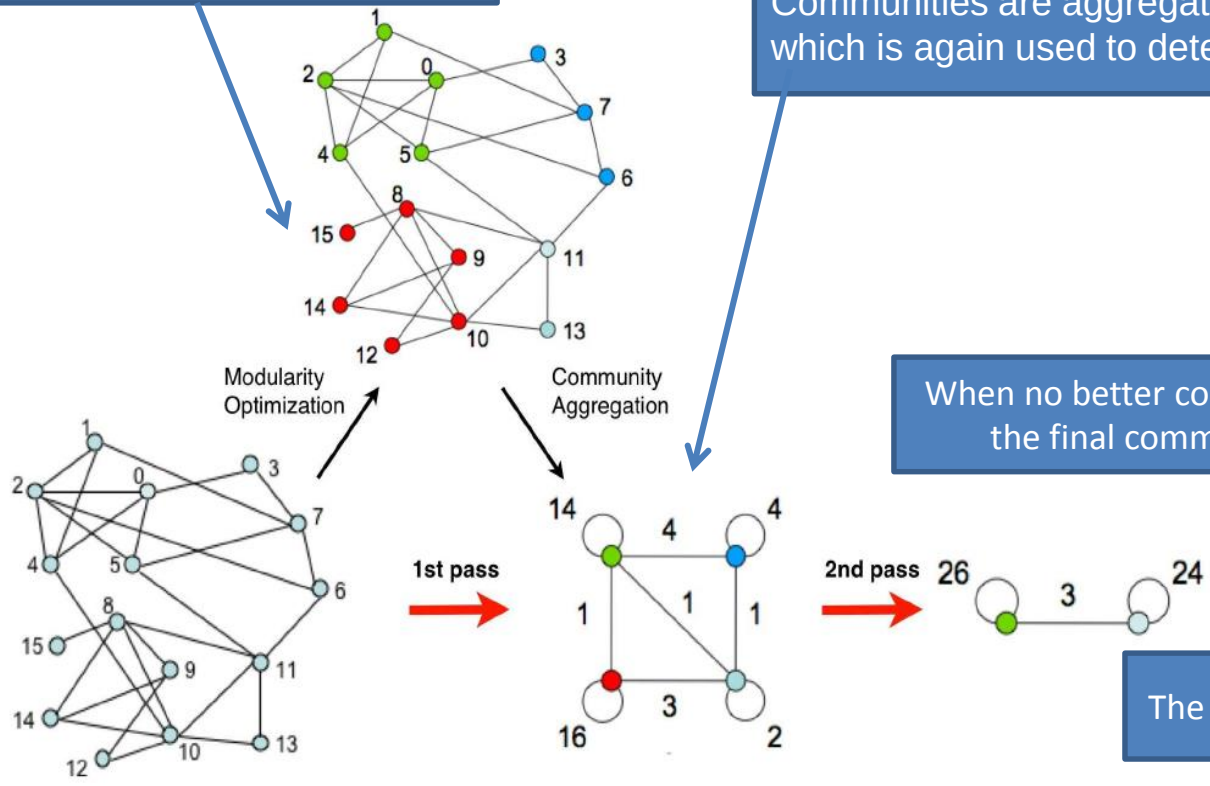
Modularity is optimized in a complex multi-level optimization scheme

First pass of optimization produces a first community structure

Communities are aggregated into a new network, which is again used to detect communities

When no better communities can be found, the final communities are returned

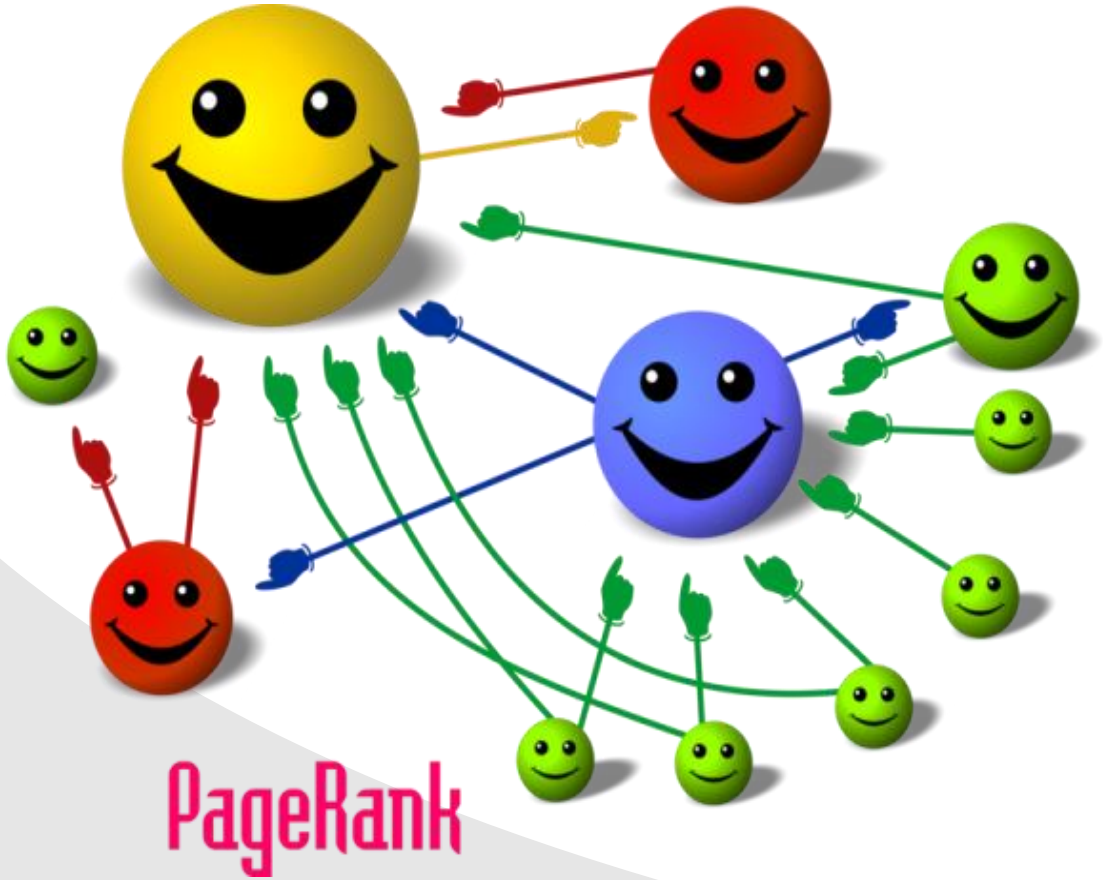
The method runs in $O(n \log n)$



[Image from original article by *Blondel et al.*, *Journal of Stat. Mech., 2008*]

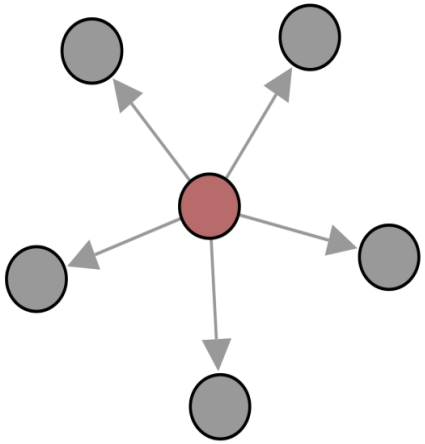
Communities discovery (4/5)
Node rankings allow to classify nodes by their level of importance in the network

e.g. PageRank allows Google to classify the pages of the Web, removing in that way the unreliant web pages from a web search

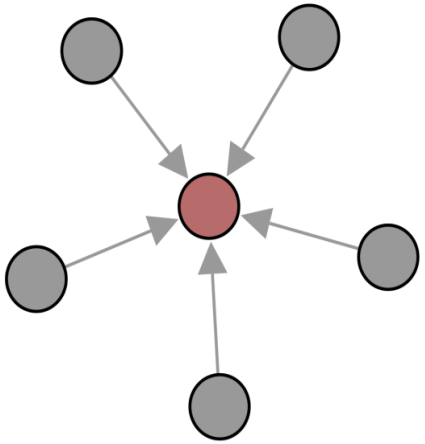


Leadership in networks may be defined following several ways depending on the application

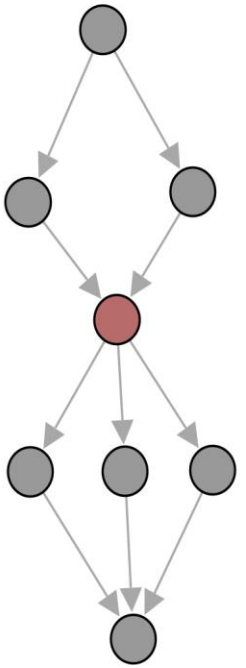
A leader may be a person who influences others



A leader may be a person often referenced by others



A leader may be a person through whom information always passes



There exist many definitions of leaders in networks, and all don't apply in all contexts

Communities discovery (5/5)

We use two ways to quantify the centrality of nodes in a network: Sum-over-all-paths-betweenness centrality and Social Leadership

Sum-over-all-paths-betweenness centrality

The centrality of a node in the network is computed as the sum of all paths going through that node, weighted by their length

$$Centrality_k = \sum_{q>0} \sum_{\substack{i,j \neq k \\ i,j \in N}} q^{-\theta} [A^q]_{ik} [A^q]_{kj}$$

Only paths are allowed, hence this sum is bounded by the number of links in the network

This measure is impossible to compute on a large network, hence we keep a good approximation by only computing it on the community of each node

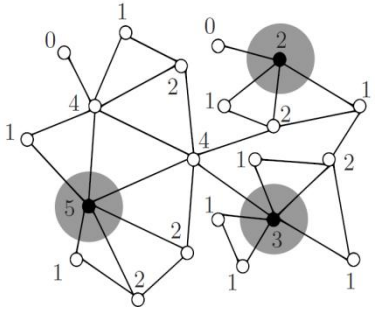
Approach seems better in theory but it hasn't been tested extensively (yet)

Social Leader (recommended)

A social leader is a node that has a higher Social Degree (SD) as his neighbors.

The social degree of a node is defined as the number of triangles made by them

$$SD_i = \sum_{j,k \in N(i)} A_{ij}$$



Ex: the social leaders are central to dense parts of the network. (image taken from C. de Kerchove's PhD thesis)

Simple technique but with a large literature (both theoric and empiric)

Agenda

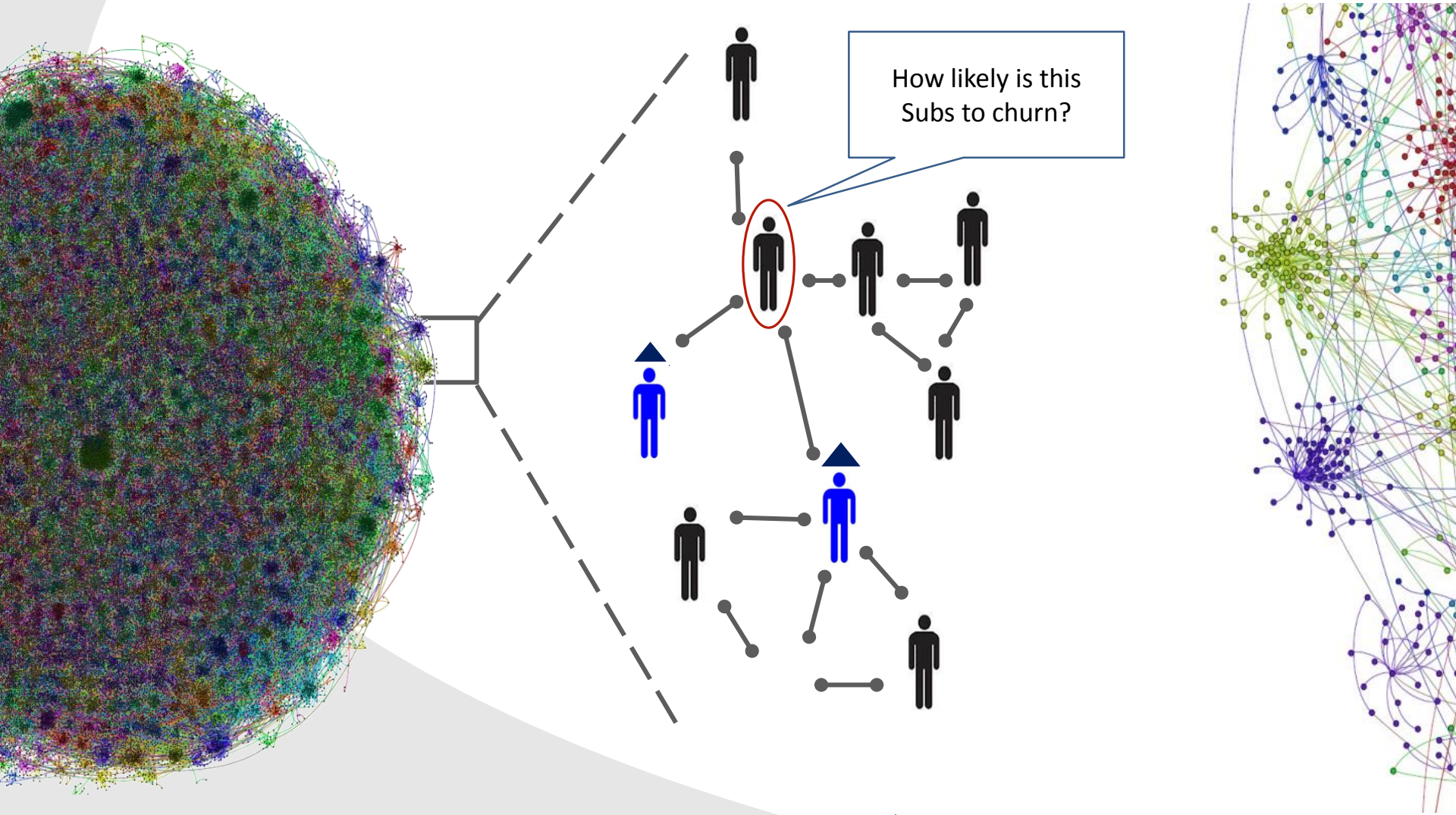
1 **Introduction to Social Network Analysis**

2 **Example: A churn prediction model**

- Context
- Methodology
- Results & improvement areas

3 Conclusion

Busines-Insight made a research collaboration on the impact of social networks* on churn prediction with a leading Telco



* We refer to Social Network Analysis (i.e. looking at communities & social graphs) rather than Social Media Analysis (i.e. crunching Facebook, Twitter & LinkedIn data)

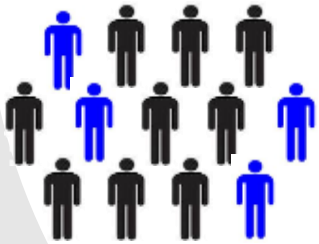
End-to-end methodology:
Churn prevention is possible thanks to data mining

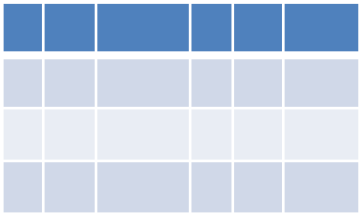
Step 1: Data Collection
Identification of current churners

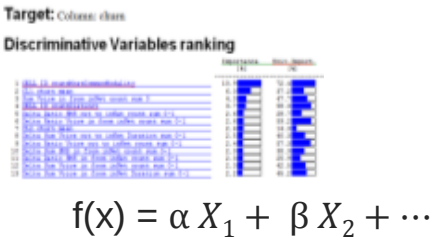
Step 2: Data Preparation
Translation of patterns into dataset

Step 3: Data mining
Extraction key variables for churn prediction

Step 4: Prediction
Scoring of potential churners







$f(x) = \alpha X_1 + \beta X_2 + \cdots$

- **Collect** only the rated Call Details Records of all the subscribers in the past 6 months
- **Flag churners** in this dataset

- **Enrich this dataset** with 3 set of variables:
  i. Socio-demographics
  ii. Usage patterns
  iii. Social network analysis

- **Feed a learning dataset** into a data mining software that will analyze it & build a prediction model with the key discriminative variables
- Following the choice of algorithm & added-variables, you will get a different formula that describes the probability to churn

- **Apply the model to a test dataset** in order to predict who the churners are
- **Compare the predictions** with the reality to measure the performance of the model (i.e. were the predictions correct?)

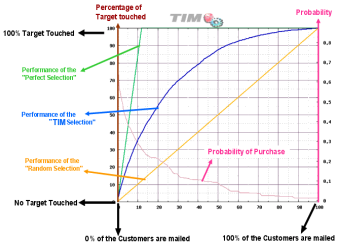
What are the numbers? Predictive model with 57% LIFT built in 6 days from raw data to scores

Step 1: Data Collection
Identification of current churners

Step 2: Data Preparation
Translation of patterns into dataset

Step 3: Data mining
Extraction key variables for churn prediction

Step 4: Prediction
Scoring of potential churners



$f(x) = \alpha X_1 + \beta X_2 + \cdots$

- Focus on **rated CDRs** only (no other data source)
- 21.4 $\bar{m}$ (million) subscribers
- 12.1 billion records
- **10 days** to copy the data

- Use of Anatella* & LinkAlytics*
- All the ETL scripts, social graph computation & subscriber behavior identification are pre-coded
- Dataset: 21.4 $\bar{m}$ rows x 657 variables (20 GB)
- **5 days** to prepare the dataset with social network variables

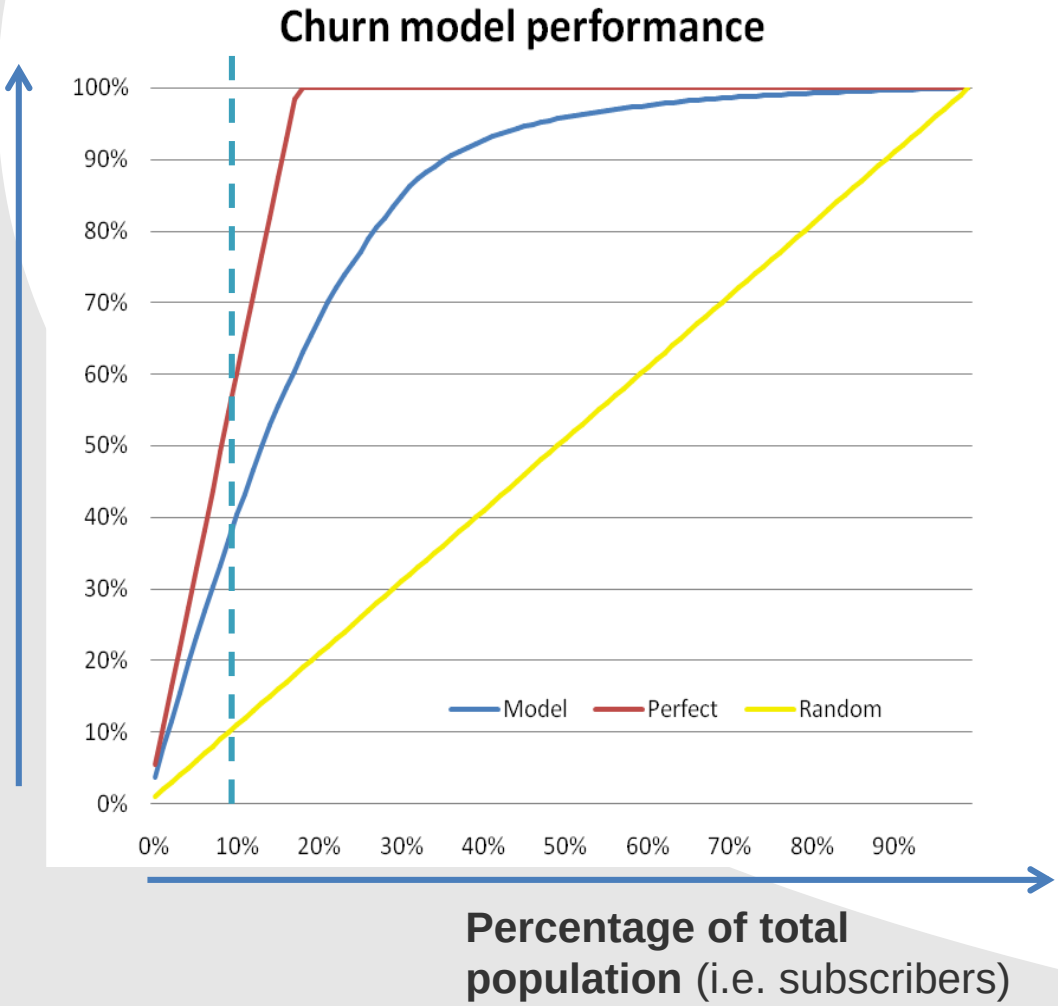
- Use of TIMi*
- Out of the 657 variables, 9 key variables are kept
- The most important variable is the % of churn within your group of friends (social variable!)
- **1 day** to extract the key variables for churn prediction

- Use of TIMi*
- **1 minute** to score the entire base with the model results

* Proprietary software

What were our results? Performance Explanation of the LIFT indicator

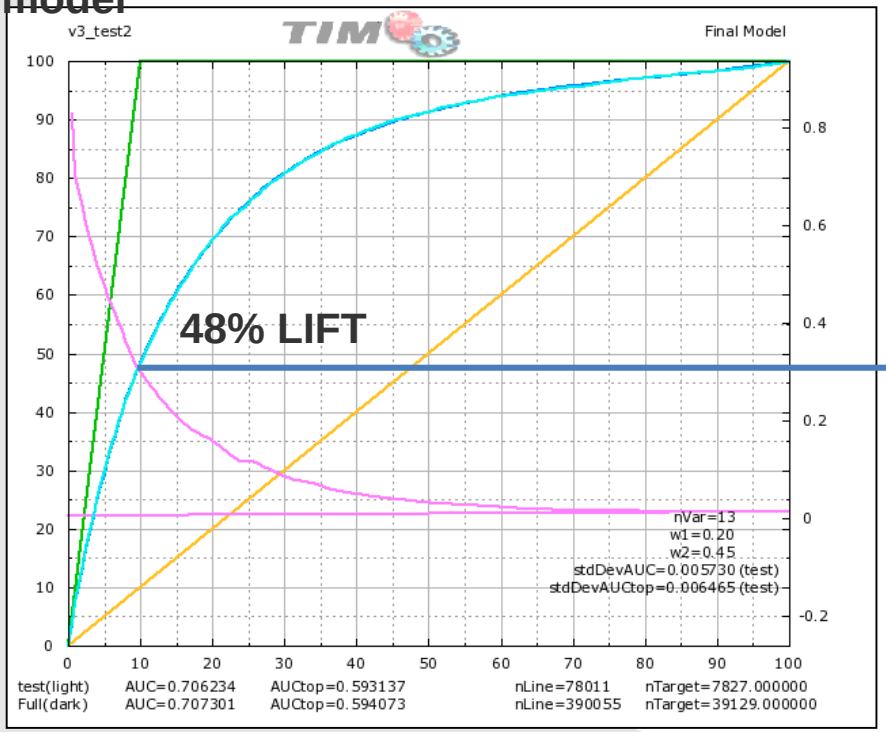
Percentage of total target identified
(i.e. churners)



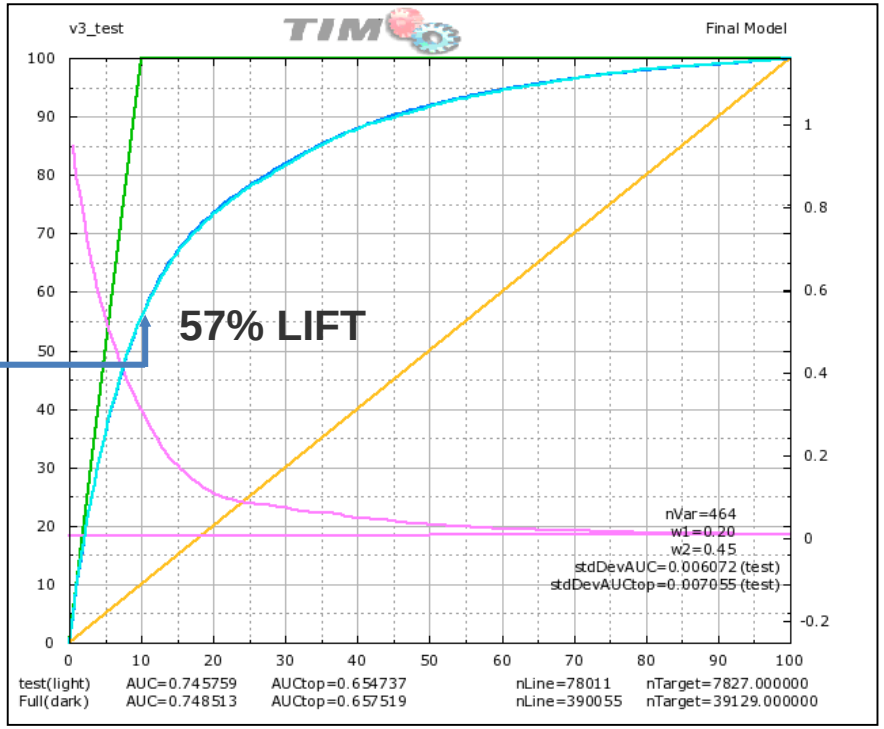
- The quality of a data mining model is how much more potential churners you reach (blue line) compared to a random selection (yellow line)
- If you send a campaign to 10% of your subs randomly, you expect to reach 10% of potential churners. With the model illustrated on the left, you reach nearly 40% of the potential churners
- With the **exact same marketing budget, you could hence reach 4 times more churners** (i.e. 40%/10%). This is the LIFT of the model

Deep-dive on LIFT results from the pilot: Social Network variables add a 9% lift (at 10%)

Recency Frequency Monetary (RFM) model



Social Network & RFM model



What were our results? Social variables (1/2)
Social interactions influence customers' decisions

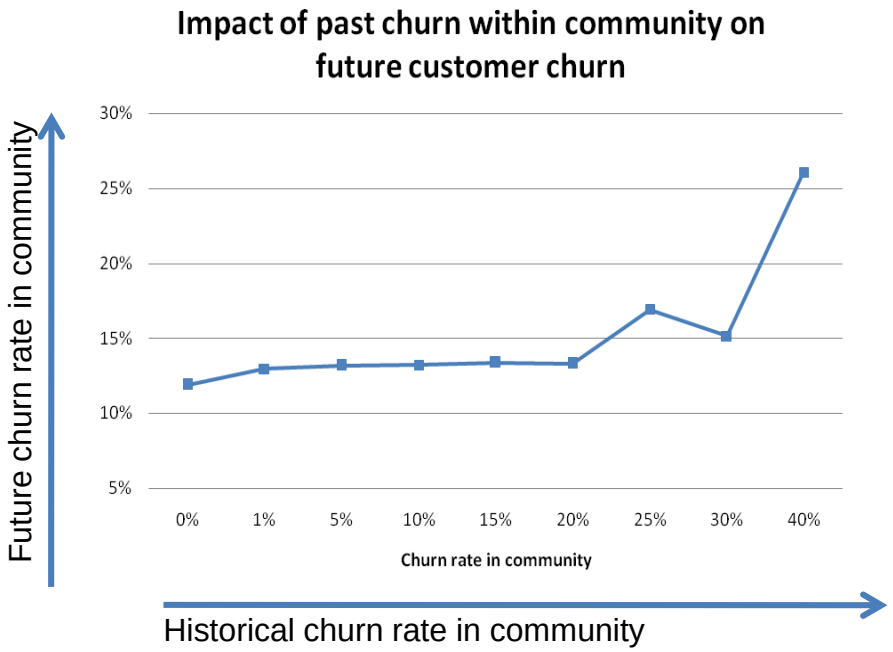
Customers who belong to a community where at least 4 customers have churned are 50% more likely to churn

Customers who belong to a community where the social leader has churned are 50% more likely to churn as well *Note: This means 1 Social leader churning in a community has the same impact as 4 regular customers churning*

On average, social leaders have twice the ARPU of other customers

Strong incentive to target the Social leaders in priority to drive retention

ILLUSTRATIVE DATA



Communities where a lot of churn happened in the past are very likely to see high churn in the future. Communities where the churn rate in the past was above 40% have future churn rates of over 25% as well.

What were our results? Social variables (2/2)
Social interactions influence customers' decisions

ILLUSTRATIVE DATA

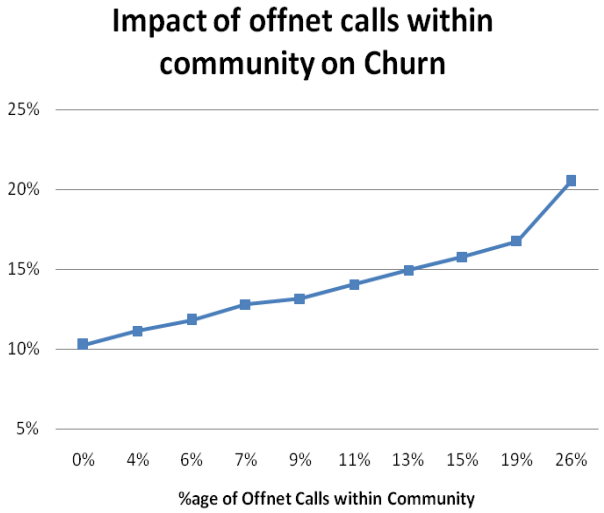
120k communities with at least 3 customers (covers 85% of the customer base)

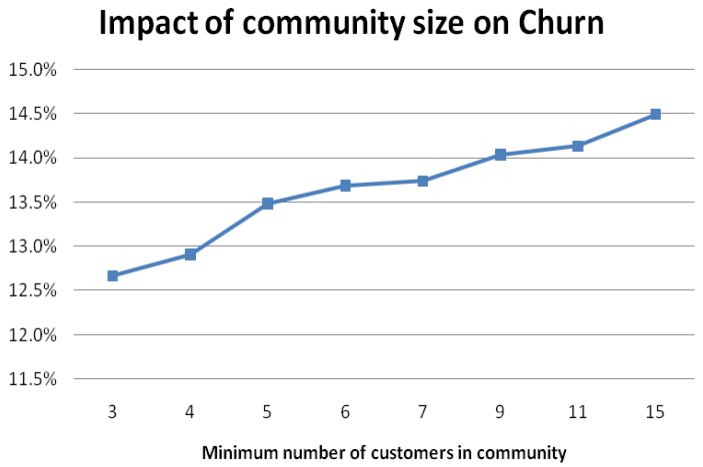
On average, communities hold 8 customers

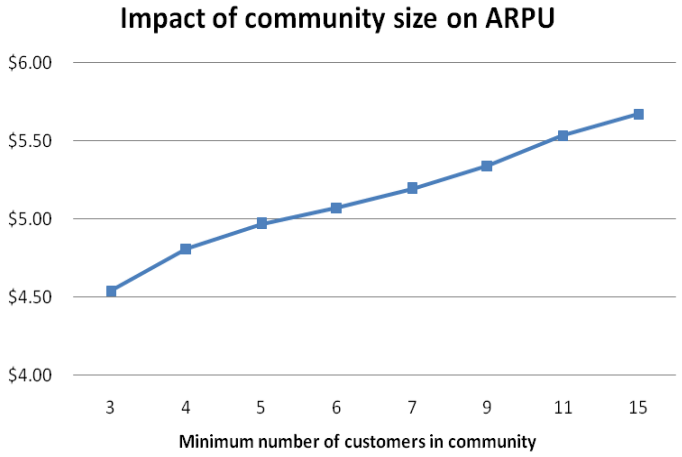
Customers who belong to a community churn 5 times less

Customers who belong to a community have an ARPU 12 times greater

Incentive for operators to foster communities (friends and families promotions)







Improvement areas for our pilot

Efficiency
- Improvement of Social Network calculations

Extend data set with new sources
- Recharge behavior: mean recharge rate
- Call Centre behavior: calls to hot-line
- Handset brand & age
- Network performance: call failure at most used cell
- Events: spends all his credit in one day
- Signaling: multi-SIM, % of active SIM

Differentiate type of churners
- Rotational churn / spinners

Integrate the data mining scores to an end-to-end campaign management process
- Create segments of customer value so that the LIFT reflects your accuracy in a segment rather than the overall population
- Create a customer profitability solution that changes the retention process based on behavior & value

Agenda

1 Introduction to Social Network Analysis

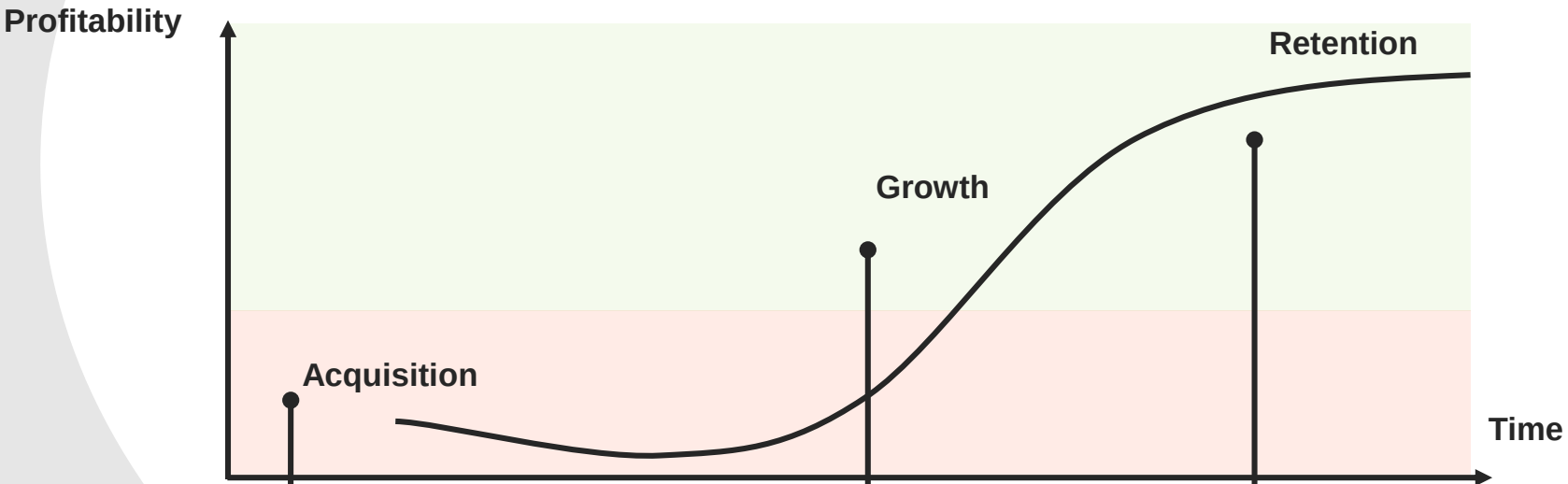
2 Introduction to Social Network Analysis

› **3 Conclusion**

Churn prediction with Social Network Analysis provided a mindset shift in how Telco's use data mining

- This pilot takes a totally different approach to several core beliefs in data mining
- **Example of challenges**
  - On which data set do you model churn?
    Entire population vs. samples
  - How long does data pre-processing take?
    Hours vs. months?
  - How do you create a data mining model?
    Automated process run from centralized data centers vs. specialized data miners in OpCo
  - Which data mining software should I use?
    Niche vs. Mainstream software
  - How often do we use data mining?
    In real-time vs. once every 6 months

SNA can be used to support all the Subscriber across their lifetime value



- Which are the Subs that should convert their spouse?
- Who are the non-Subs who would benefit the most to convert to Telco X?

- Which services can I **cross-sell** to this subscriber based on their community (e.g. Data, Mobile Money, etc.)?
- Who is most likely to adopt this **new product**?

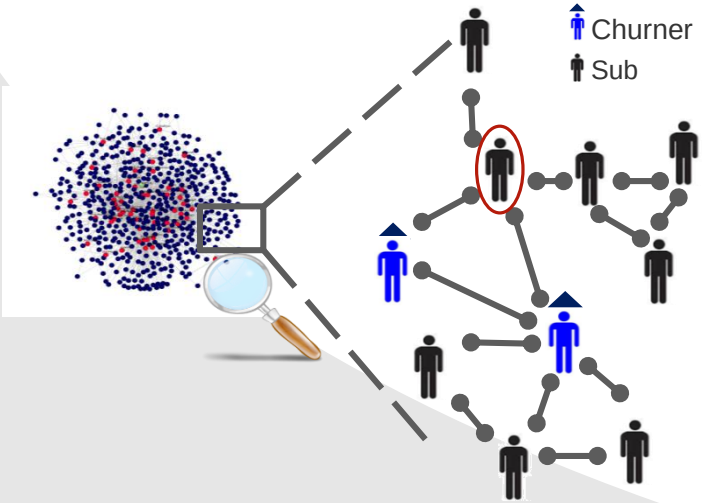
- Who is likely to **churn?** Which friend is most likely to retain them?

[Case study #1] Improve churn prediction using SNA metrics

Case study
Consumer
Churn

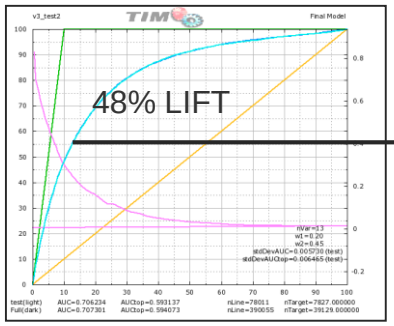
Context

- **Core assumption:** our group of friends influences our decision to churn
- **Telco X SA & Real Impact** have started a research pilot on this topic
- Case involved **21,4 million Telco X subs &** their behavior over 6 months

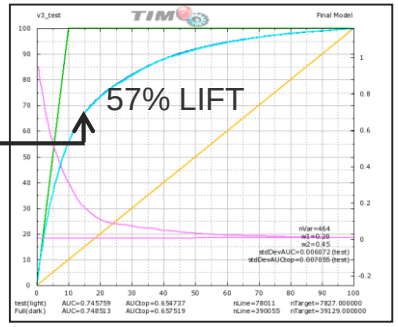


Results

- SNA model **increased the churn model accuracy** with a prediction of 9% additional potential churners
- Based on the campaign success & « win-back » rate, potential savings of 9% more subscriber lifetime value



Recency Frequency Quantity model



RFQ with SNA metrics model

[Case study #2] Increase cross-sell campaign effectiveness by launching a viral marketing campaign on social leaders

Case study
Consumer
Cross-sell

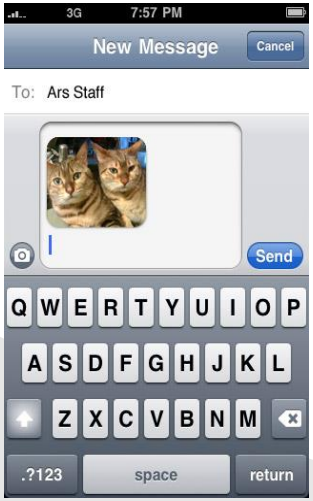
Context

- **Objective:** increase MMS usage
- Use **SNA metrics** to **identify social leaders -** subscribers that influence others
- **Metrics of success:**
  - ✓ Response rate = % of Subs accepting the offer
  - ✓ Viral spread = # referrals per Subs targeted

Results

- **Targets** are prepaid and hybrid Social Leaders with on-net spending lower than Rx per month
- **Offer** 100 free MMS per month during 3 months + more if a friend registers to the offer
- **Results:**
  - ✓ response rate = 5x better (35% vs. 7%)
  - ✓ viral spread = 2x better (10% vs. 5%)

Increase adoption of MMS by targeting social leaders with a special offer

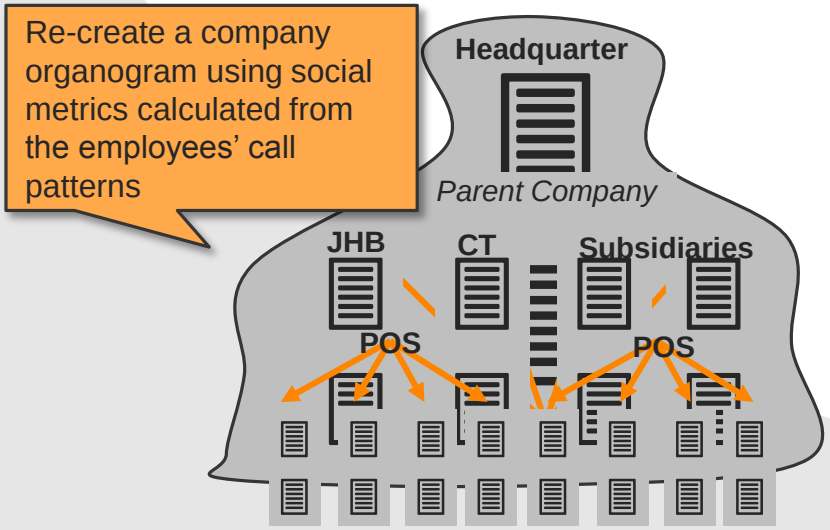


[Case study #3] Increase sales force effectiveness by identifying decision makers in business accounts

Case study
Business
Sales

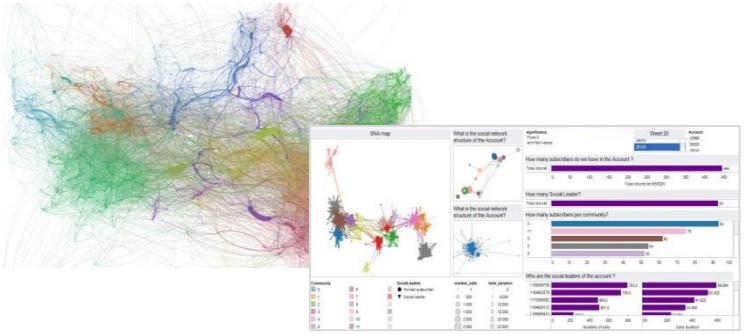
Context

- **Objective:** Define a model using SNA metrics to identify the decision makers in a business account based on CDRs
- **Metrics of success:** Identify at least 60% of all the decision makers contained in a validation list
- **Received data set:** Receive a data set with 3 months of rated CDRs for the 962 employee of Company X & a validation list of ~300 high profiles



Results

- Manage to identify up to **80% of the decision makers** in the Company X account based on limited CDRs analysis
- 80% of the members of each community are actually part of the same department



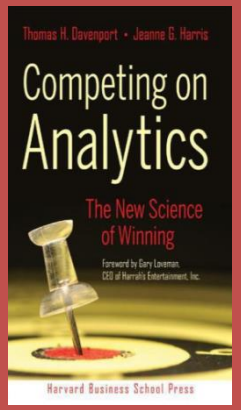
Variables in model

Leadership measure	3 variables
Communication out of account	3 variables
Communication in account	2 variables
Centrality of position in network	1 variable

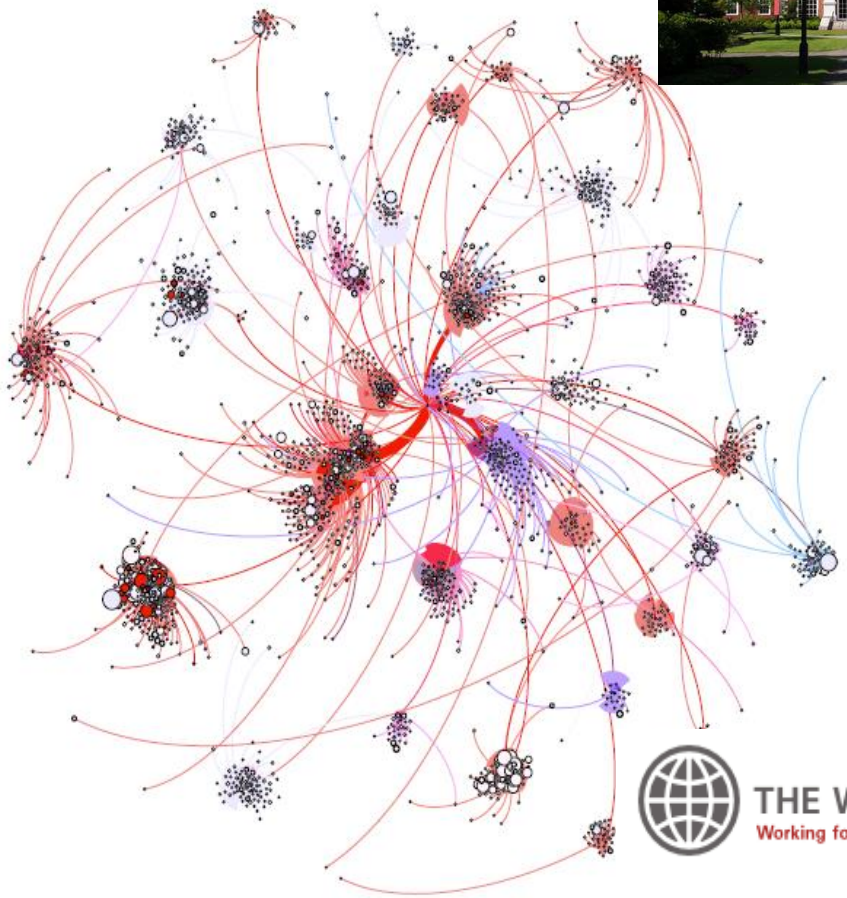
Future deployments : Tom Davenport and the World Bank want to write about our results

Tom Davenport

- Author: **“Competing on Analytics”**
- Prof. Harvard Business School
- Senior advisor at **Deloitte Analytics**
- Member of **SAP** Innovation Council
- Advisory Board member **Real Impact Analytics**



*Currently scoping an article for **Harvard Business Review and the World Bank***



THE WORLD BANK
Working for a World Free of Poverty

Thanks for your Attention

For more information, please consult our website:

http://www.business-insight.com