KU LEUVEN

GOTCHA!
IMPROVING FRAUD DETECTION TECHNIQUES USING SOCIAL NETWORK ANALYSIS

Prof. Dr. Bart Baesens
Véronique Van Vlasselaer

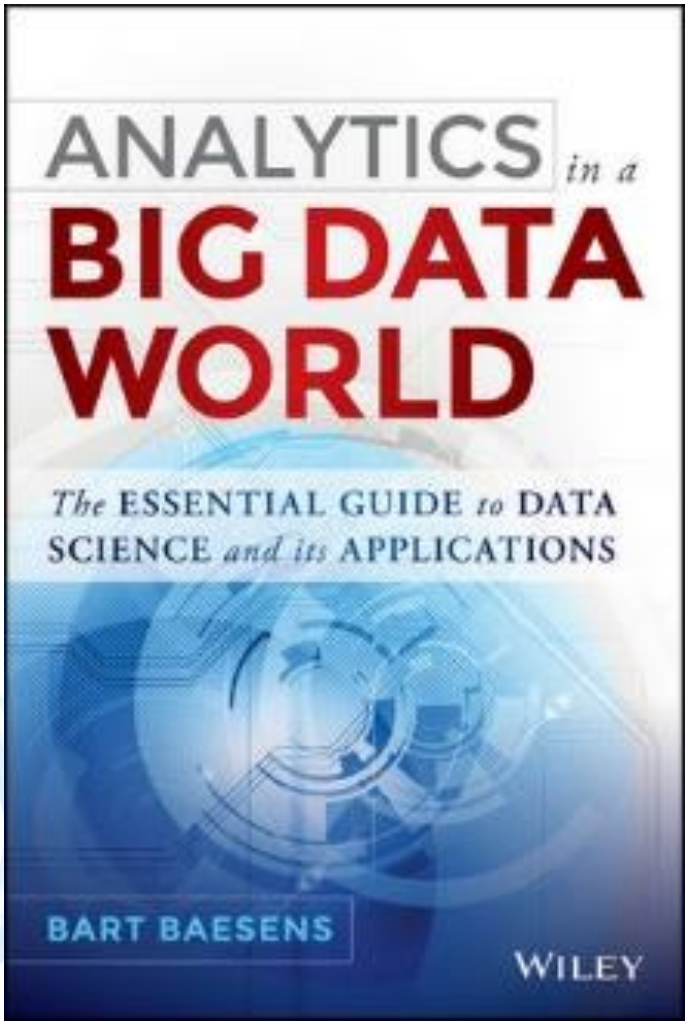
SAS Forum 2014, Louvain-la-Neuve

Presenter: Bart Baesens

- Studied at the Catholic University of Leuven (Belgium)
  - Business Engineer in Management Informatics, 1998
  - PhD. in Applied Economic Sciences, 2003
- PhD. Title: Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
- Associate professor at the KU Leuven, Belgium
- Associate professor at Vlerick Leuven Ghent Management School
- Lecturer at the School of Management at the University of Southampton, United Kingdom
- Research: analytics, credit risk, fraud, marketing, …
- Youtube/Facebook/Twitter: DataMiningApps
- www.dataminingapps.com
- Bart.Baesens@kuleuven.be

KU LEUVEN

Book


ANALYTICS in a
BIG DATA
WORLD
The ESSENTIAL GUIDE to DATA
SCIENCE and its APPLICATIONS
BART BAESENS
WILEY

Presenter: Véronique Van Vlasselaer

- Studied at the Catholic University of Leuven (Belgium)
  - Information Systems Engineer, 2012
- PhD student @ KU Leuven, department of “Decision Sciences and Information Management”
- PhD title: FAIR: Forecasting and Network Analytics for Management of Payment Risk
- Fields of expertise: Social Network Analysis, Fraud Detection, Net lift modeling, …

- Contact: Veronique.VanVlasselaer@kuleuven.be
- Website: www.dataminingapps.com

KU LEUVEN

Outline

- Fraud detection

- (Social) network analysis
  - Networks and components
  - Is fraud a social phenomenon?

- GOTCHA!
  - Social security fraud
  - Credit card fraud

Fraud Detection

KU LEUVEN

Fraud?

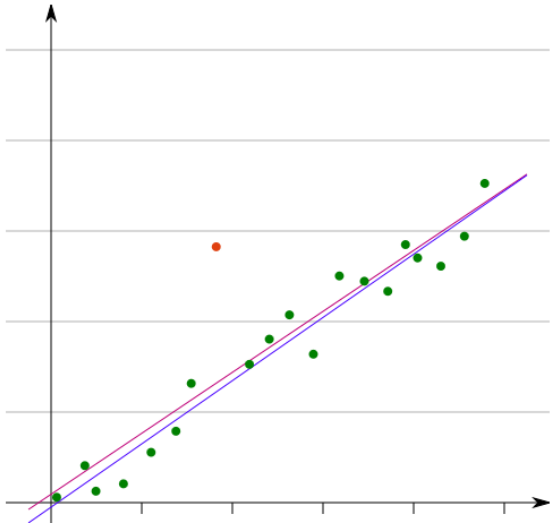
- Present in many critical human processes:
  - Credit card transactions
  - Insurance claim fraud
  - Opinion fraud
  - Social security fraud
  - Call behavior fraud
  - Scientific fraud
  - Terrorism
  - …

Fraud?

- Anomalous behavior
  - *Outlier detection*: abnormal behavior and/or characteristics in a data set might often indicate that that person perpetrates suspicious activities.

    - Data set level

      Behavior of a person/instance does not comply with overall behavior.
      E.g., illegal set up of customer account



source: metaoptimize.com

Fraud?

- Anomalous behavior
  - *Outlier detection*: abnormal behavior and/or characteristics in a data set might often indicate that that person perpetrates suspicious activities.

    - Data item level

      Behavior of a person/instance does not comply with normal behavior of that person/instance. E.g., identity theft

Date & Time	Day	Duration	Origin	Destination	Fraud
1/01/95 10:05:01	Mon	13 mins	Brooklyn, NY	Stamford,CT	
1/05/95 14:53:27	Fri	5 mins	Brooklyn, NY	Greenwich,CT	
1/08/95 09:42:01	Mon	3 mins	Bronx, NY	White Plains, NY	
1/08/95 15:01:24	Mon	9 mins	Brooklyn, NY	Brooklyn, NY	
1/09/95 15:06:09	Tue	5 mins	Manhattan, NY	Stamford, CT	
1/09/95 16:28:50	Tue	53 sec	Brooklyn, NY	Brooklyn, NY	
1/10/95 01:45:36	Wed	35 sec	Boston, MA	Chelsea, MA	BANDIT
1/10/95 01:46:29	Wed	34 sec	Boston, MA	Yonkers, NY	BANDIT
1/10/95 01:50:54	Wed	39 sec	Boston, MA	Chelsea, MA	BANDIT
1/10/95 11:23:28	Wed	24 sec	White Plains, NY	Congers, NY	
1/11/95 22:00:28	Thu	37 sec	Boston, MA	East Boston, MA	BANDIT
1/11/95 22:04:01	Thu	37 sec	Boston, MA	East Boston, MA	BANDIT

source: Fawcett and Provost, 1997

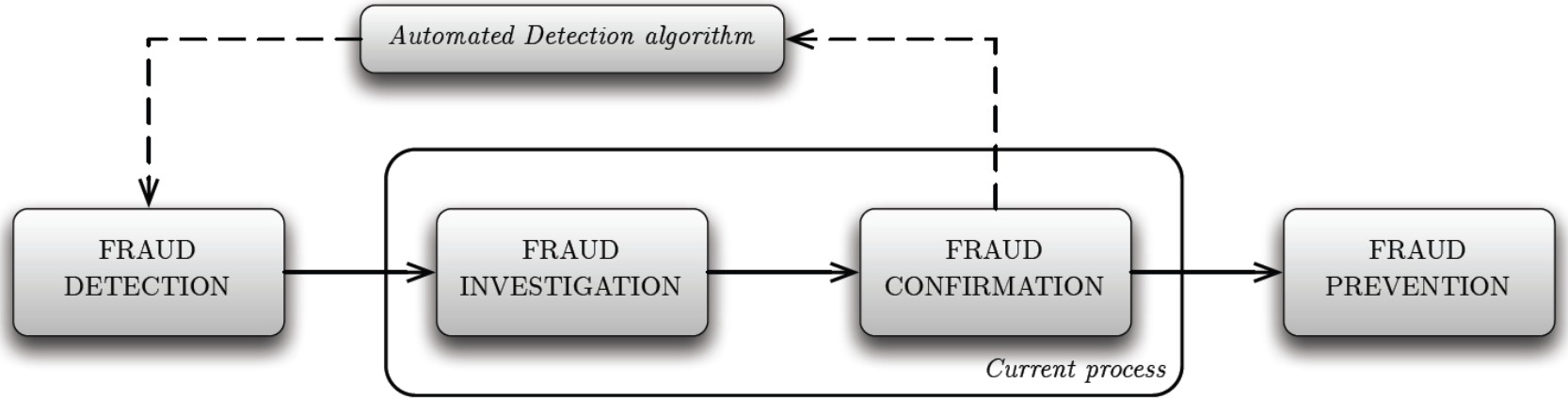
Fraud?

- Normal behavior
  - How to detect mr. Hyde?

KU LEUVEN

Fraud?

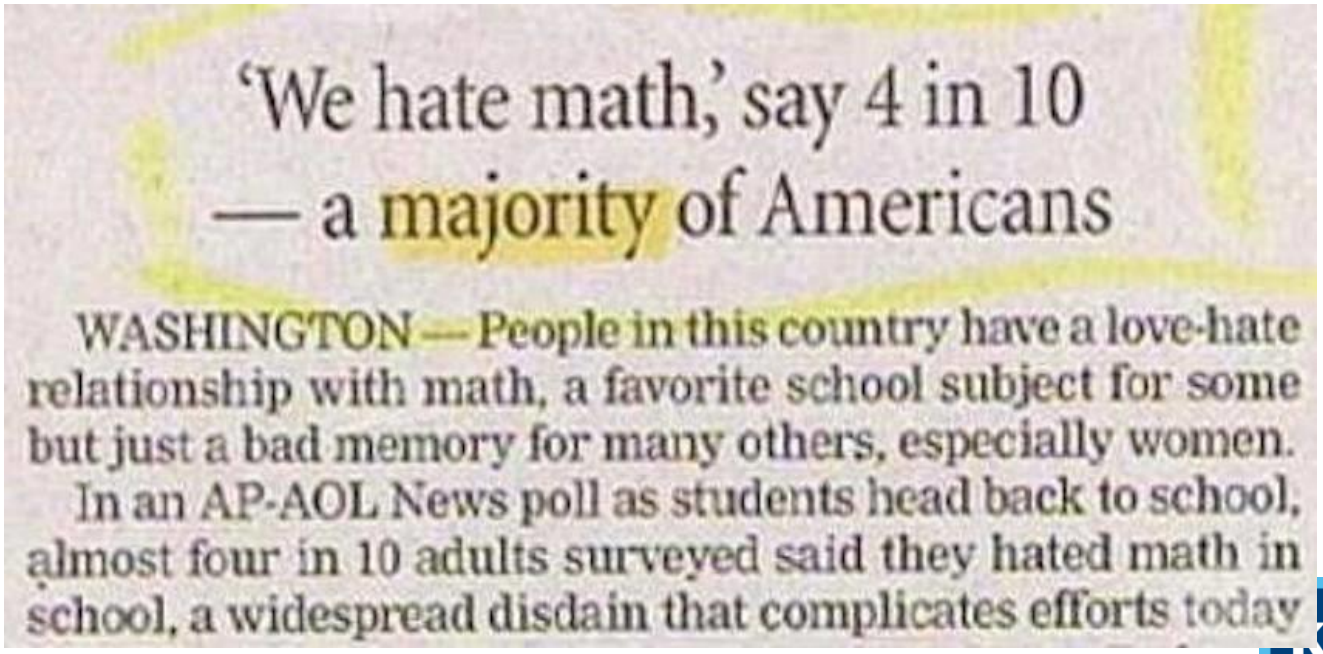
- Normal behavior
  - Tendency of fraudsters to blend into the surroundings (camouflage)
  - No longer fraud by opportunity, but carefully planned
  - Need for new techniques:
    - Evolving from *descriptive statistics* towards *predictive statistics* (data mining)
    - Learning from historical data to judge upon new observations,
    - Detection of *patterns* that initially seem to comply to normal behavior, but in reality instigate fraudulent activities.



KU LEUVEN

Properties of fraud detection models

- Accuracy (AUC, precision and recall)
- Operational efficiency (e.g. 6 second rule in credit card fraud)
- Economical cost
- Interpretability (i.e. make sense)



'We hate math,' say 4 in 10 — a majority of Americans

WASHINGTON — People in this country have a love-hate relationship with math, a favorite school subject for some but just a bad memory for many others, especially women.

In an AP-AOL News poll as students head back to school, almost four in 10 adults surveyed said they hated math in school, a widespread disdain that complicates efforts today

Fraud Detection

- Challenges:

Fraud is an *uncommon*, *well-considered*, *imperceptibly concealed*, *time-evolving* and often *carefully organized* crime which appears in many types and forms.

Fraud Detection

- Challenges:

Fraud is an *uncommon*, *well-considered*, *imperceptibly concealed*, *time-evolving* and often *carefully organized* crime which appears in many types and forms.

Uncommon:

- extremely skewed class distribution
- Big data, but only few fraudulent observations (often < 1%)
- Hard for data mining algorithms to learn from
- Rebalancing techniques: oversampling, undersampling, *SMOTE*

KU LEUVEN

Fraud Detection

- Challenges:

Fraud is an *uncommon*, *well-considered*, *imperceptibly concealed*, *time-evolving* and often *carefully organized* crime which appears in many types and forms.

Well-considered:

- Complex fraud structures are carefully planned
- Outlier detection no longer sufficient: combination of patterns, preferably well-hidden
- Historical changes in behavior: *temporal weighting*

Fraud Detection

- Challenges:

Fraud is an *uncommon*, *well-considered*, *imperceptibly concealed*, *time-evolving* and often *carefully organized* crime which appears in many types and forms.

Imperceptibly concealed

- Subtlety of fraud: imitating normal behavior, even in identify theft
- Fraudsters are often first "*sleeping*", pretending to be a good customer

Fraud Detection

- Challenges:

Fraud is an *uncommon*, *well-considered*, *imperceptibly concealed*, *time-evolving* and often *carefully organized* crime which appears in many types and forms.

Time-evolving

- Fraudsters learn from their mistakes and those of their predecessors (Jensen, 1997)
- How does past fraud affects the present?
- *Time-dependent* fraud probability

Fraud Detection

- Challenges:

Fraud is an *uncommon*, *well-considered*, *imperceptibly concealed*, *time-evolving* and often *carefully organized* crime which appears in many types and forms.

Carefully organized

- Relationships between fraudsters:
  - Fraudulent transactions often occur at the same (set of) merchants
  - Companies often inherit many suspicious resources from past fraudulent companies (social security fraud)
  - Fraudsters tend to call/contact the same (set of) people once they stole someone's identity

Social Network Analysis (SNA)

Is fraud a social phenomenon?

KU LEUVEN

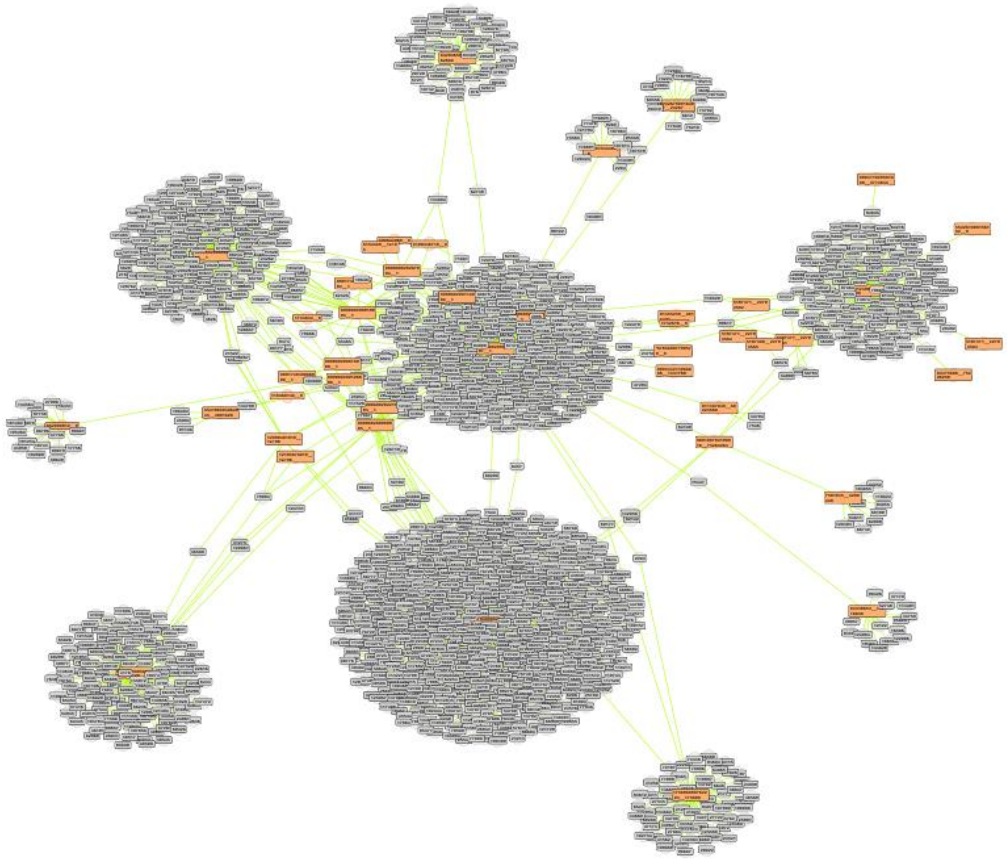
Social Network Analytics

KU LEUVEN

Social Network Analytics

Social Network Analysis (SNA):

Deriving useful patterns and insights by exploiting the relational structure between objects.



KU LEUVEN

Social Network Analytics: Components

Nodes:

The objects of the network.

- People
- Computers
- Reviewers
- Companies
- Credit card holders
- …

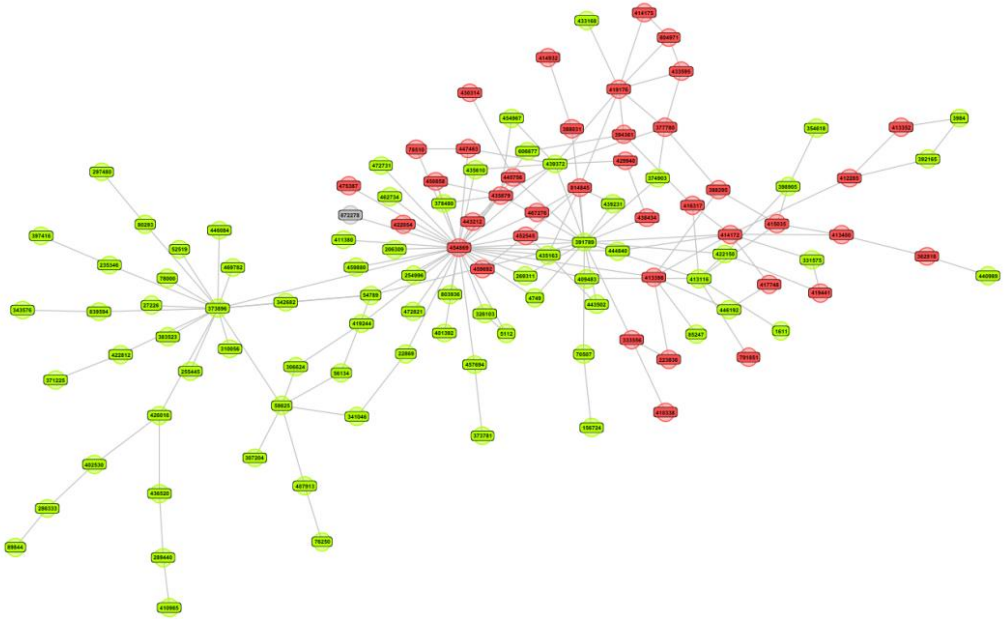
Links:

The relationships between objects

- Call record
- File sharing
- Product reviews
- Shared suppliers/buyers
- Merchant
- ….

Is fraud a social phenomenon?

Homophily in social networks (from sociology):
People have a strong tendency to associate with other whom they perceive as being similar to themselves in some way.
e.g.: same city, hobbies, interests…

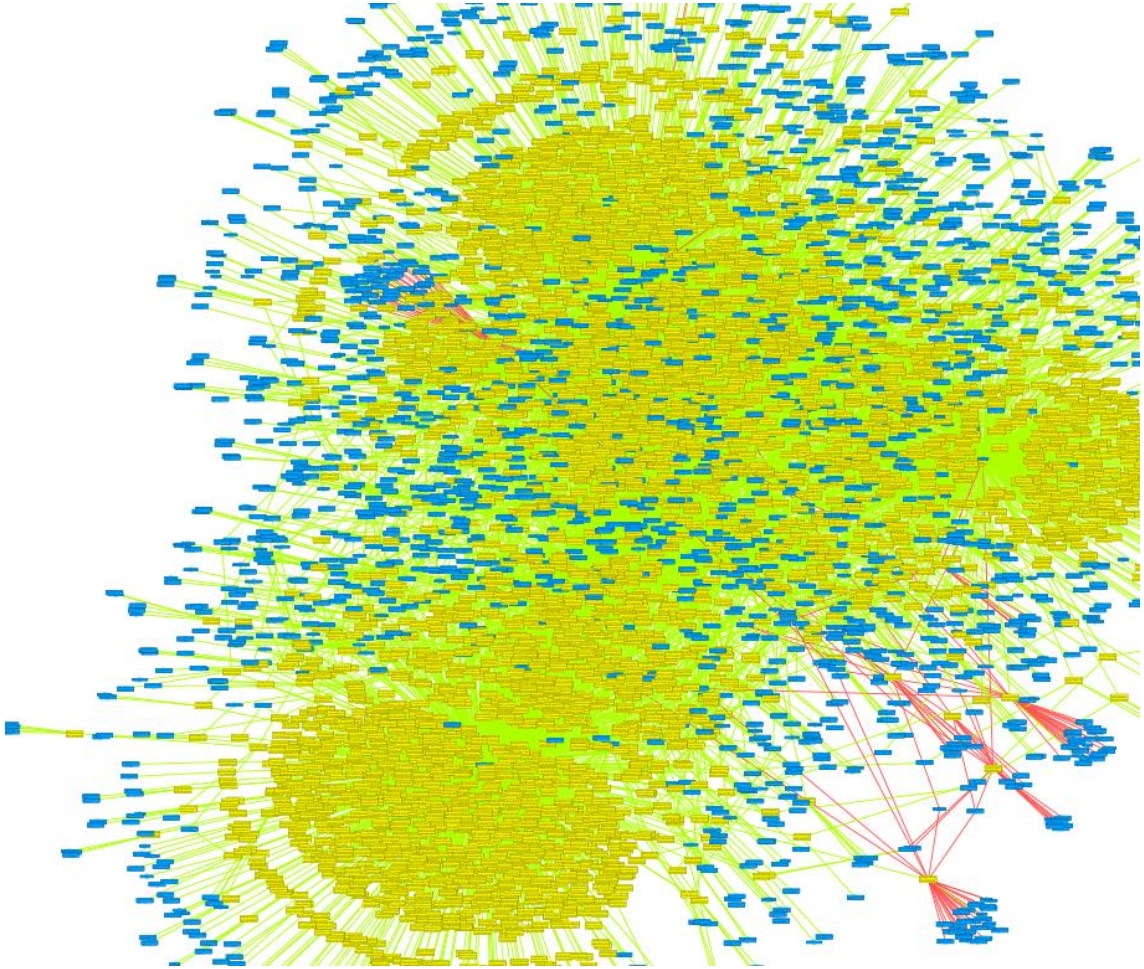


Question: Does the network contain statistically significant patterns of homophily?

KU LEUVEN

Is fraud a social phenomenon?

Credit card transaction fraud:

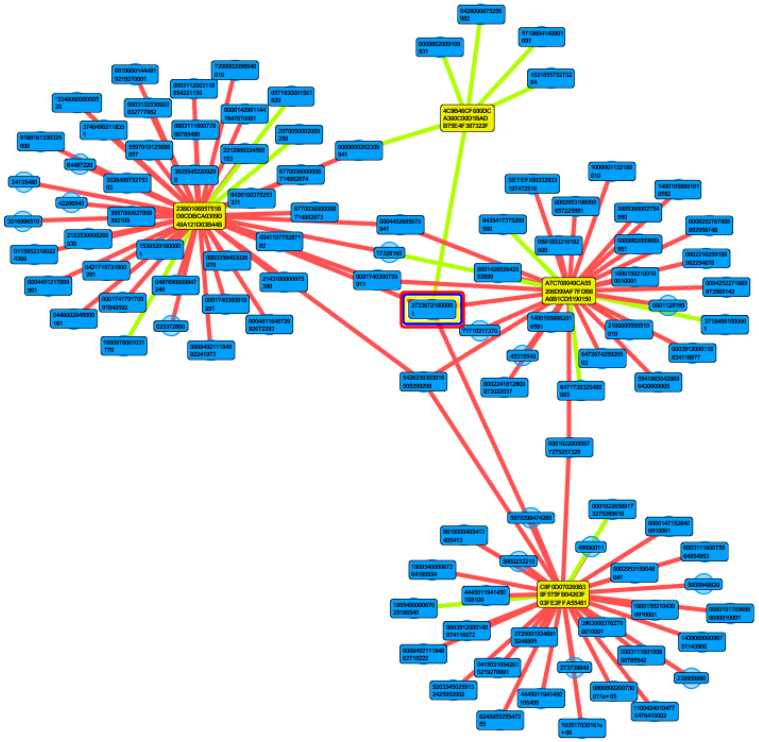


KU LEUVEN

Is fraud a social phenomenon?

Credit card transaction fraud:

- Stolen credit cards (yellow nodes) are often used in the same stores (blue nodes)
- Store itself also processes *legitimate* transactions to cover their fraudulent activities

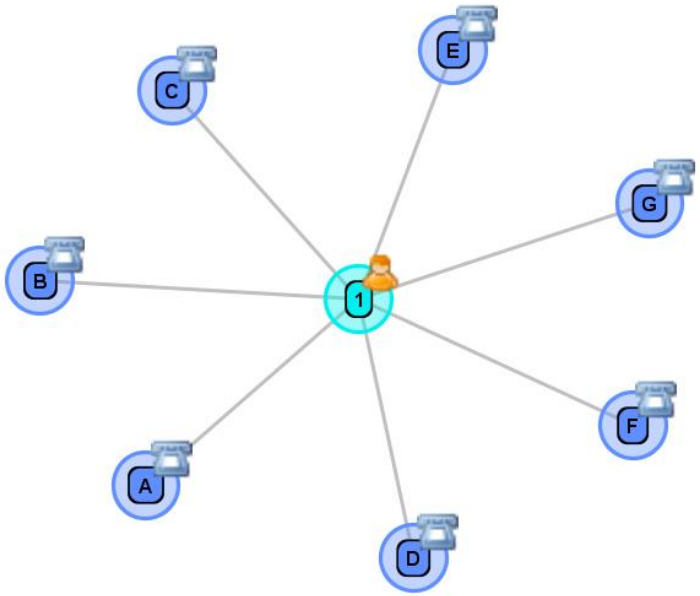


KU LEUVEN

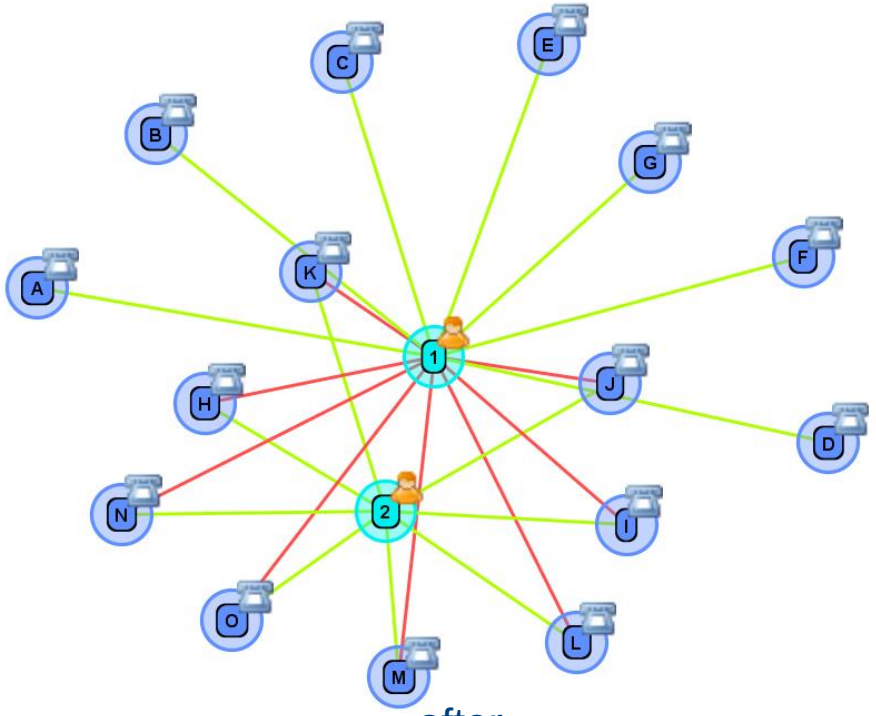
Is fraud a social phenomenon?

Identify theft:

- Before: person calls his/her frequent contacts
- After: person also calls new contacts which *coincidentally* overlap with another persons contacts.



before



after

KU LEUVEN

Is fraud a social phenomenon?

Social security fraud:

- Companies are frequently associated with other companies that perpetrate suspicious/fraudulent activities.

KU LEUVEN

Is fraud a social phenomenon?

- Analysis of fraudulent networks?
  - Fraudsters appear to be closely related to/have many things in common with other fraudsters
  - How can we include network information in detection tools?

    Networks are no default *data representations*:

    - Visualization
      - Visualizing network and manually assess which instances need to be passed to further investigation
    - Link analysis
      - Linking different data sources. Often this results in a large database with instance-specific, instead of aggregated data
    - Network analysis
      - *Featurization process*: extracting features for each network object based on its neigbhorhood.

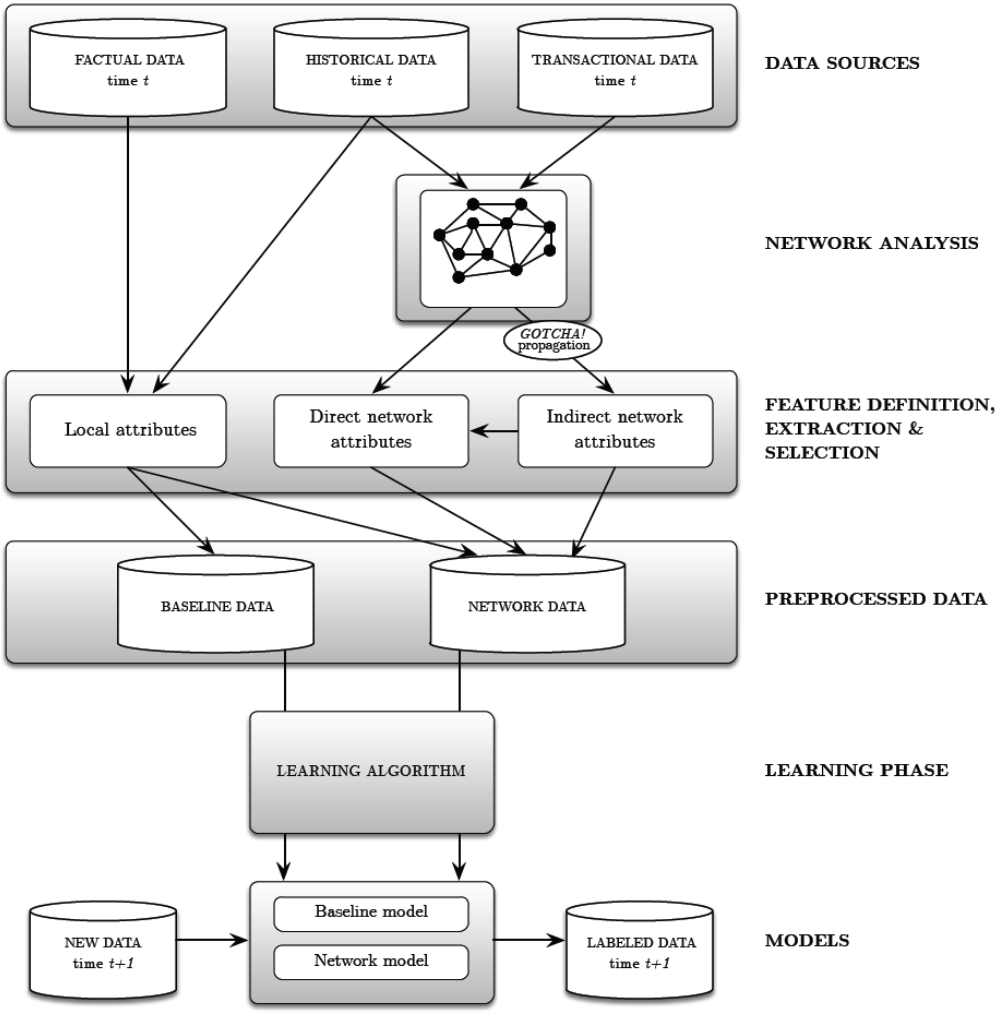
GOTCHA!
A NEW FRAUD DETECTION APPROACH

Overview

KU LEUVEN

Gotcha! Fraud Detection Tool

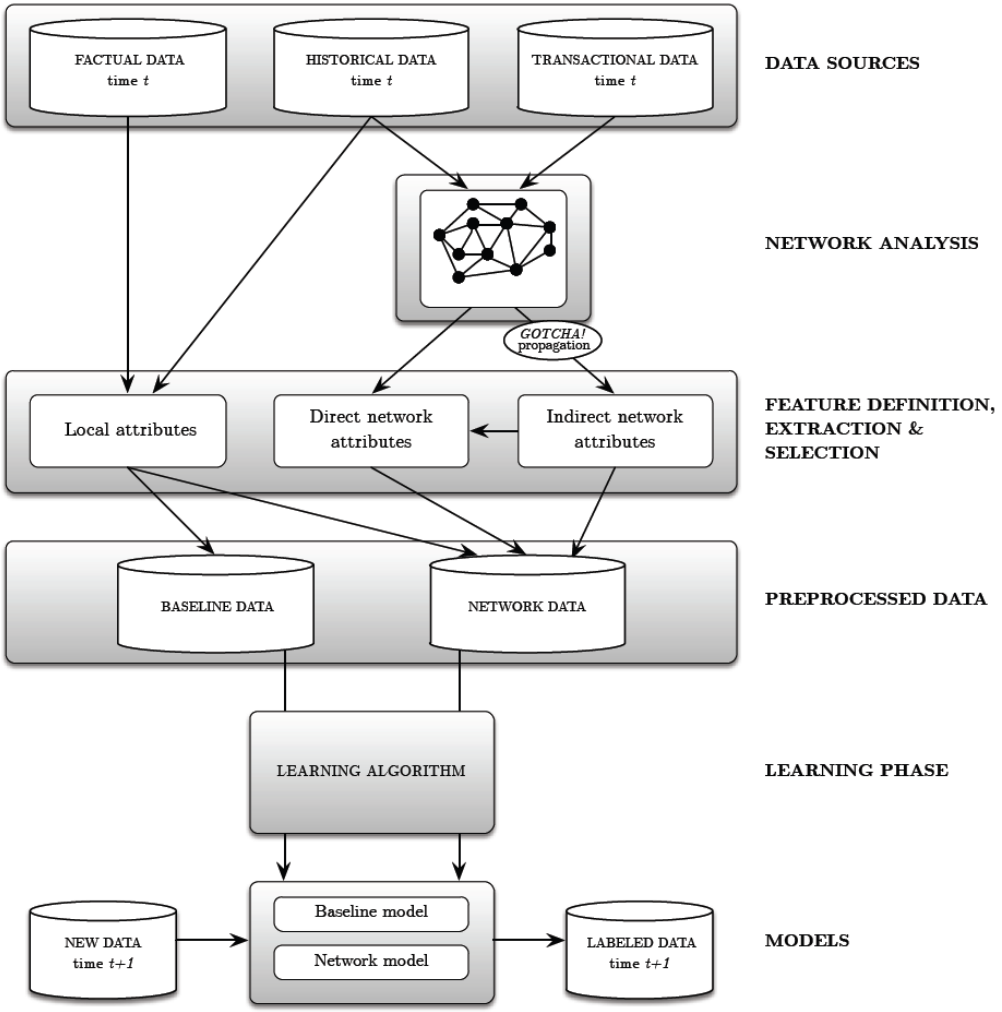
- Gotcha!:
  - Detection approach which integrates both intrinsic and network-related features
  - Start from three types of data sources: factual data, historical data (past relationships and changes in behavior), and transactional data (current relationships between instances)
  - Network analysis part: both deriving first-order neighborhood as well as global network-specific characteristics for each instance



KU LEUVEN

Gotcha! Fraud Detection Tool

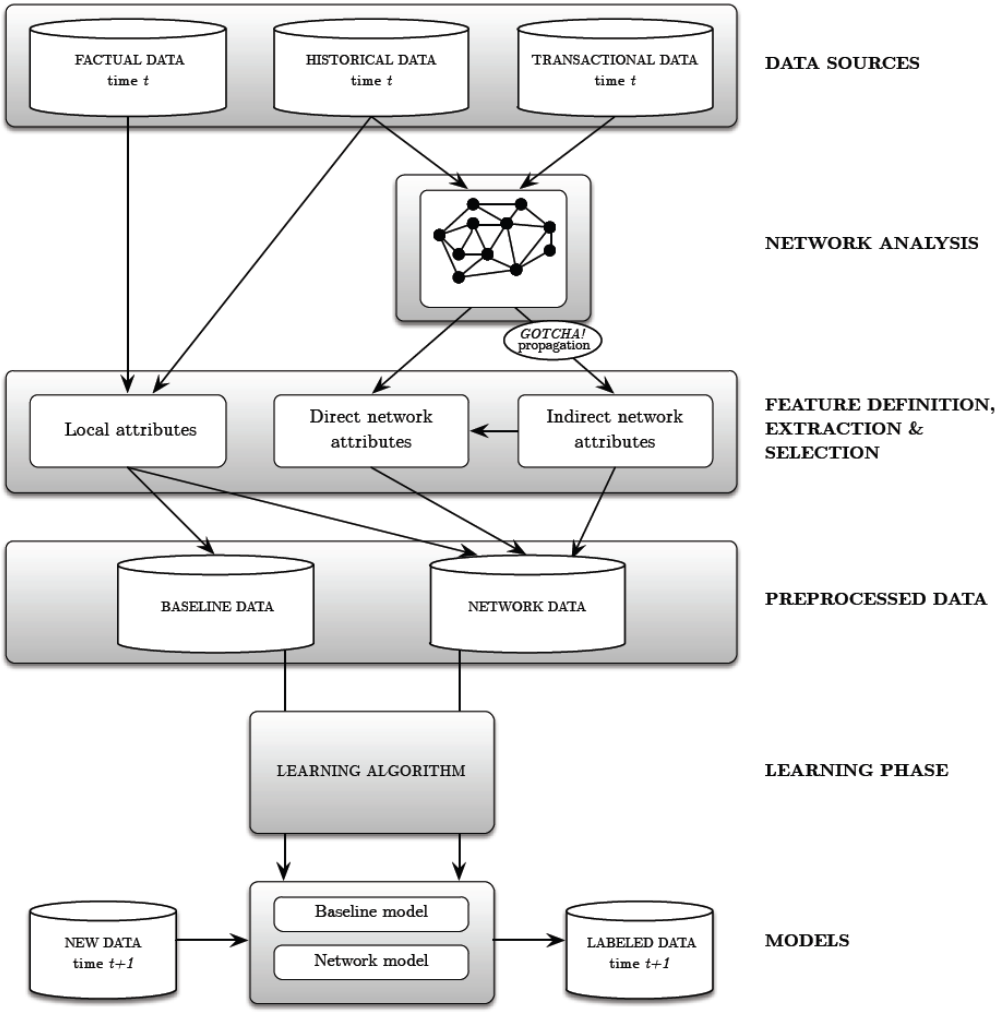
- Gotcha!:
  - *Carefully organized*: Do the interrelations between instances play an important role in the diffusion of fraud?
  - *Time-evolving*: Does the effect of fraud diminishes over time? Does the influence of (past) relationships between instances diminish over time?
  - *Imperceptibly concealed*: What is the effect of few fraudulent instances in the network? How do they affect the other network objects?



KU LEUVEN

Gotcha! Fraud Detection Tool

- Gotcha!:
  - *Well-considered*: integration of intrinsic, direct and indirect network attributes.
  - *Uncommon*: only less than 1% of all the instances is fraud. How can we emphasize fraud to guide learning algorithms?



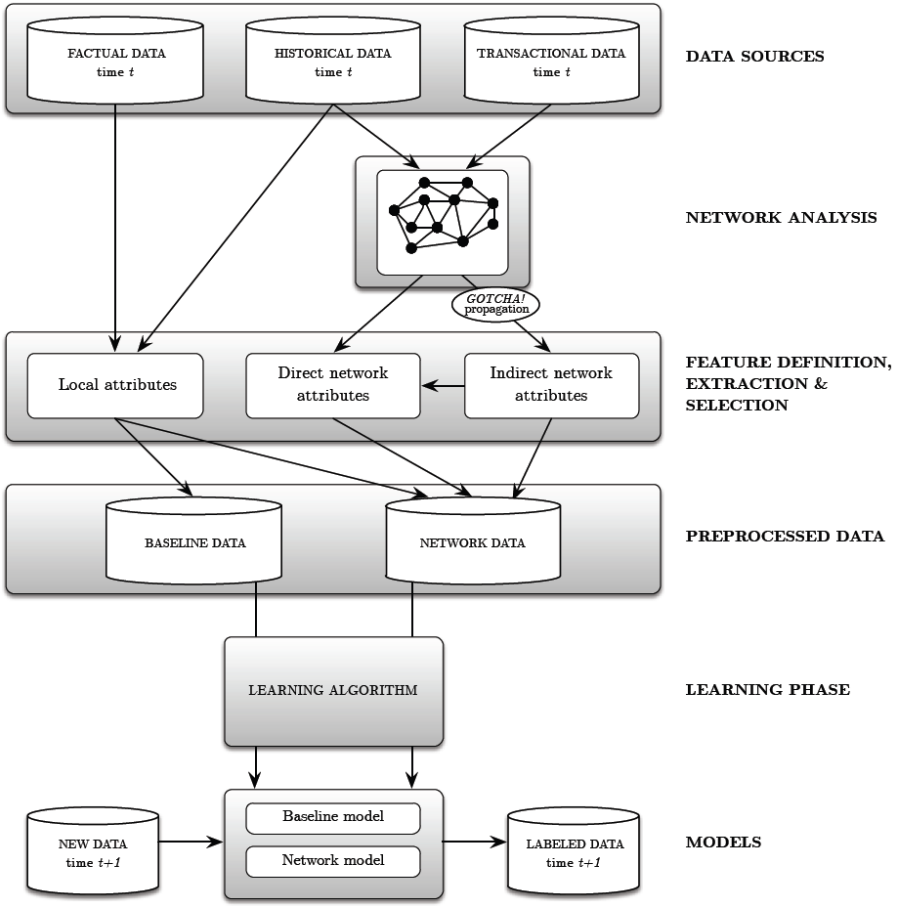
KU LEUVEN

Gotcha! Fraud Detection Tool

- Two applications:
  - Social security fraud
    - Companies are related to each other by means of *shared resources*
    - Some companies *intentionally* do not contribute to the government, and are bankrupt
    - *Goal*: find those companies that form a high-risk of perpetrating fraud in the future?
  - Credit card transaction fraud
    - People are related to stores where they make their purchases
    - Some credit cards are used in illicit transactions (stolen/copied).
    - *Goal*: find those transactions that are likely to be illicit?

Gotcha! Outline

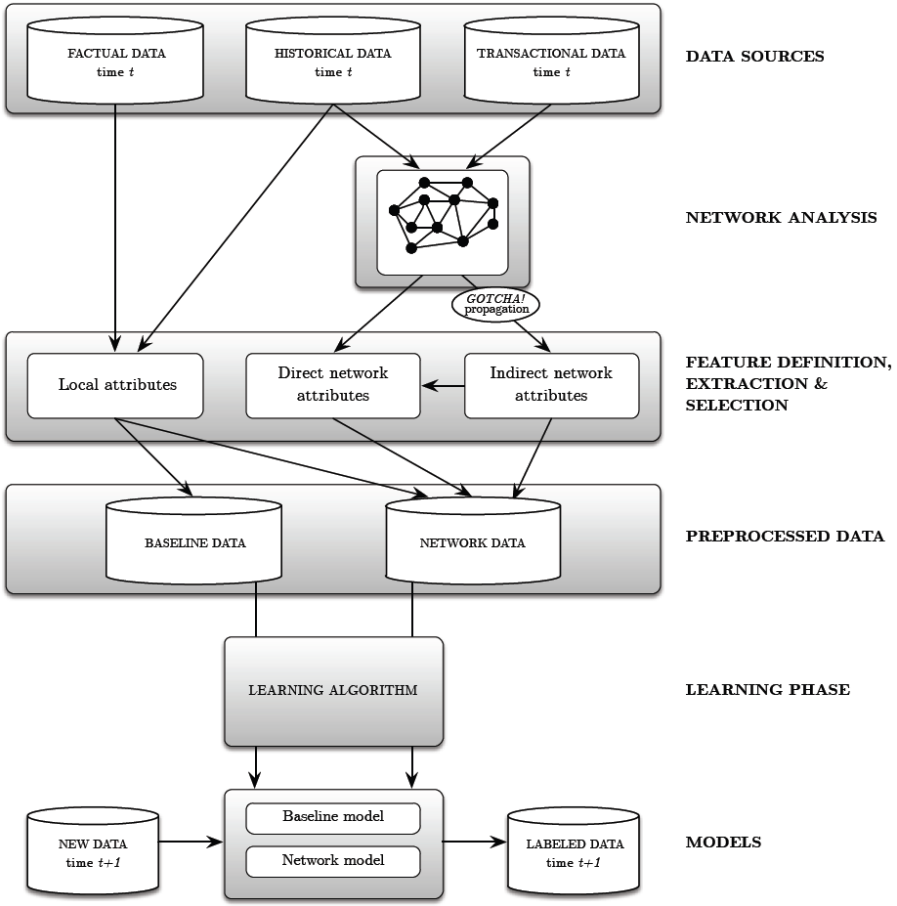
1. Data sources
2. Network analysis
3. Feature definition, extraction and selection
4. Learning phase
5. Detection and prevention model



KU LEUVEN

Gotcha! Outline

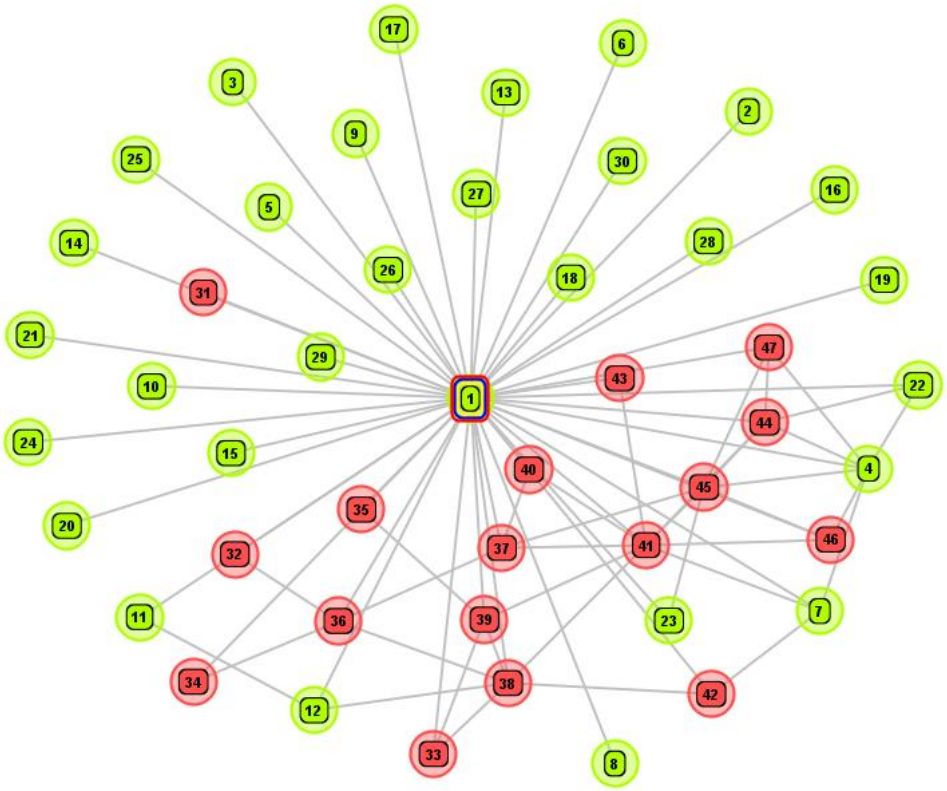
1. Data sources
2. Network analysis
3. Feature definition, extraction and selection
4. Learning phase
5. Detection and prevention model



KU LEUVEN

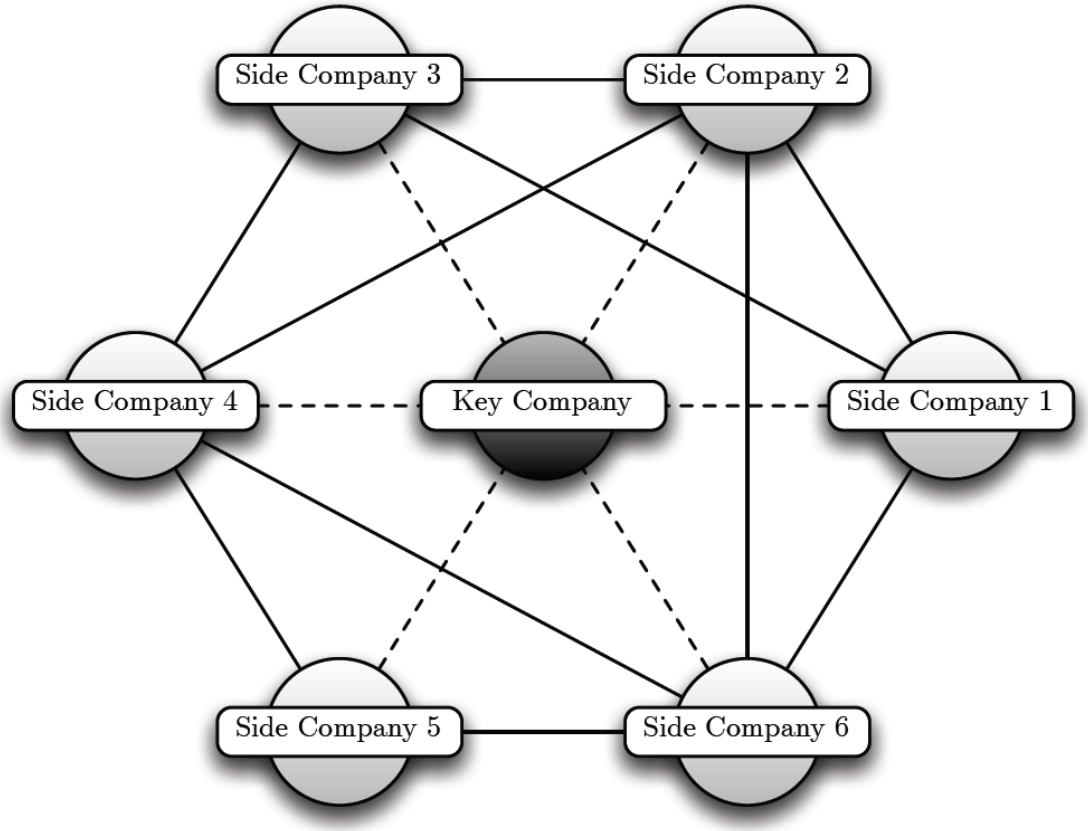
Gotcha! Network representation

- Network representation:



KU LEUVEN

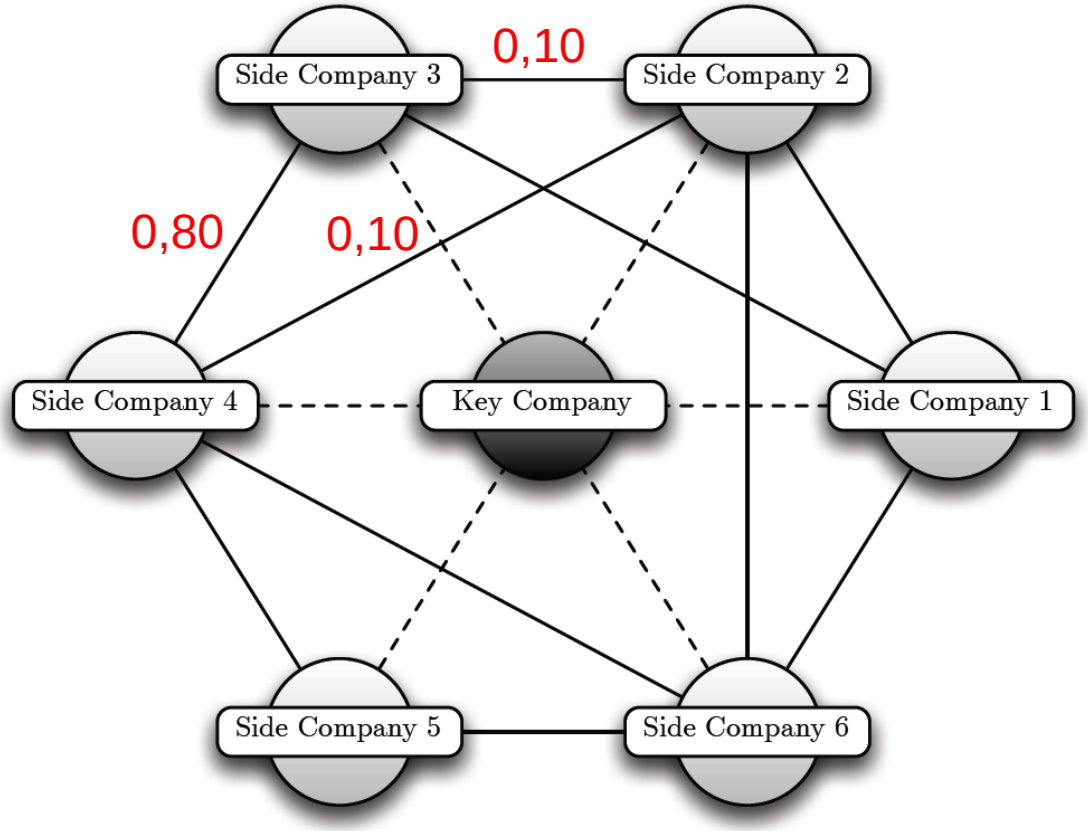
Gotcha! Network representation



KU LEUVEN

Gotcha! Network representation

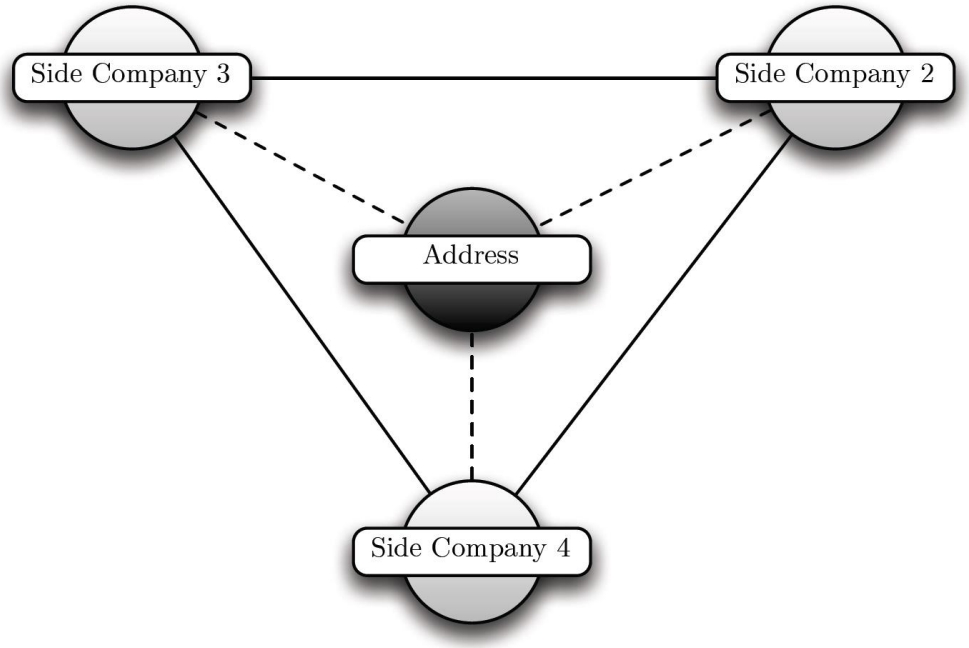
Link weight?



KU LEUVEN

Gotcha! Network representation

Link weight?



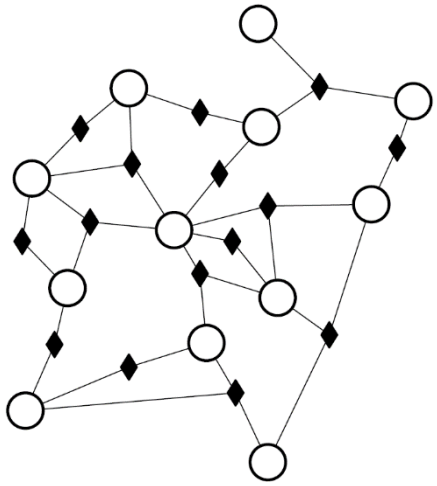
KU LEUVEN

Gotcha! Network representation

In most real-life networks, different types of objects are related to each other

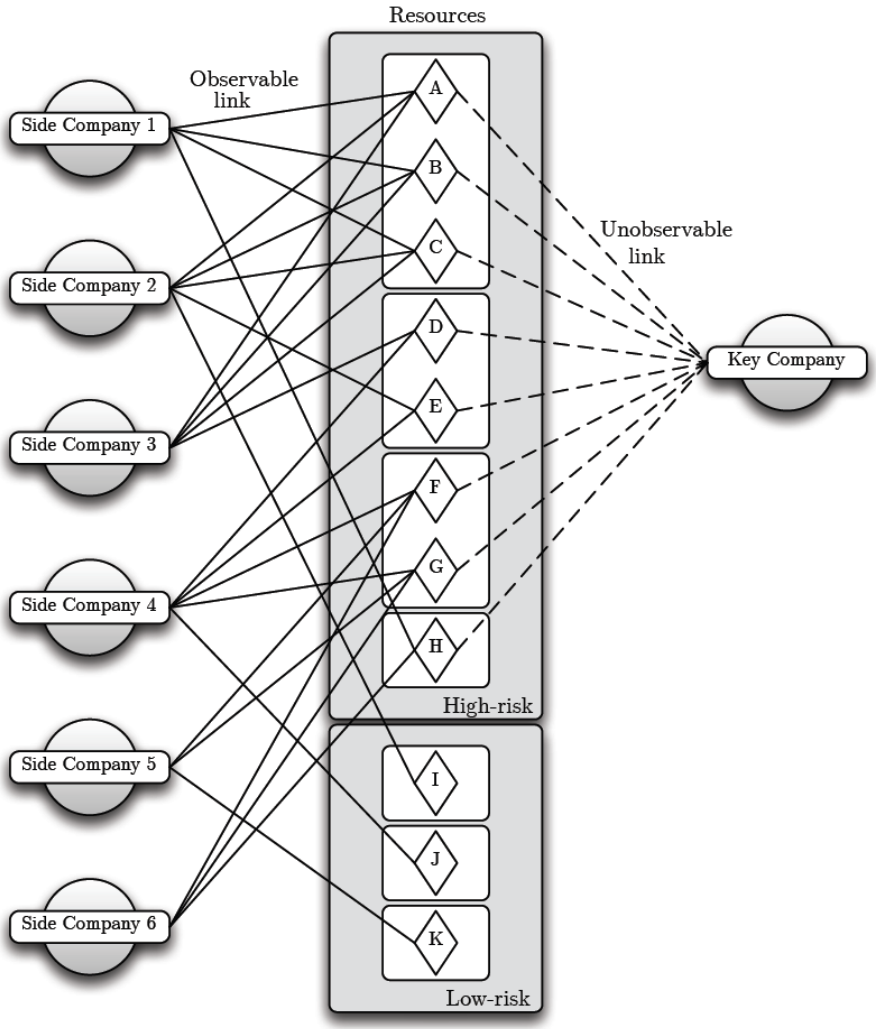
e.g.

reviewers	-	review
people	-	(fraudulent) activities
credit card	-	merchant
people	-	insurance claims
companies	-	resources



Introduction of bipartite graphs or bigraphs

Gotcha! Bigraphs

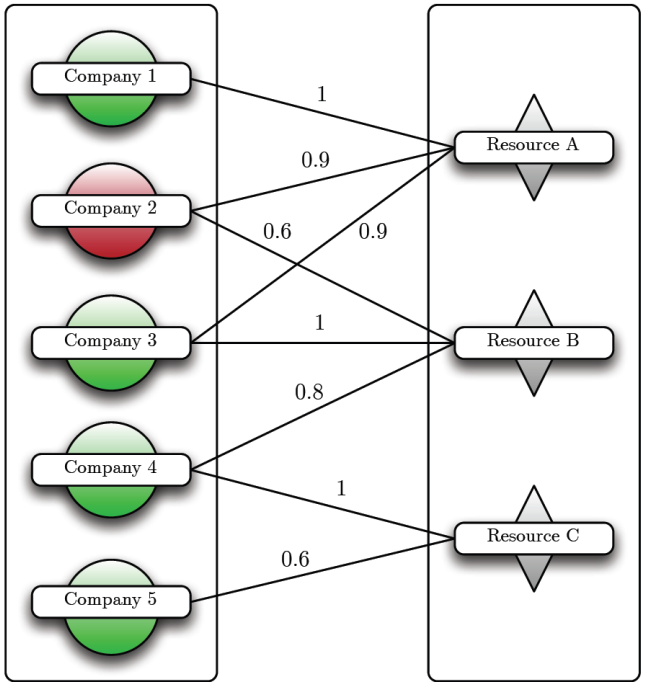


KU LEUVEN

Gotcha! Bigraphs

Link weight?

- Recency of the relationship
- Possiblity to include past relationships:
  - ~0: old relationship
  - 1: current relationship



KU LEUVEN

Gotcha! Propagation algorithm

Goal: Derive an exposure score for every network object

Questions:

1. How can we start from *only* confirmed fraud objects to infer an *exposure score* for the other network objects?

   i.e. label the *legitimate* companies / transactions

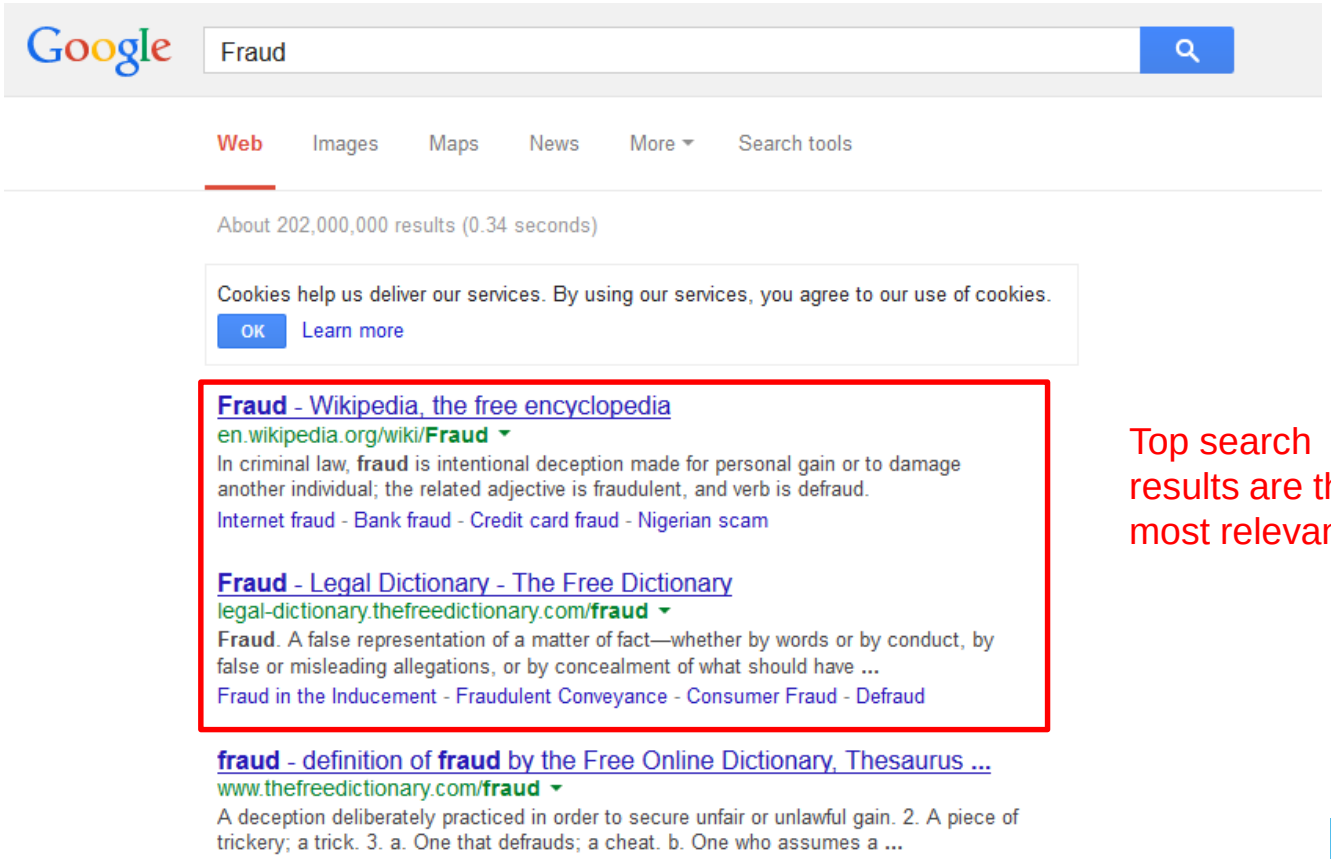
2. How can we use evidence from one type of node to infer an *exposure score* to another type of node

   i.e. label the resources / merchants and credit card holders

KU LEUVEN

Gotcha! Propagation algorithm

- Solution:
  - Google PageRank algorithm: extension to fraud

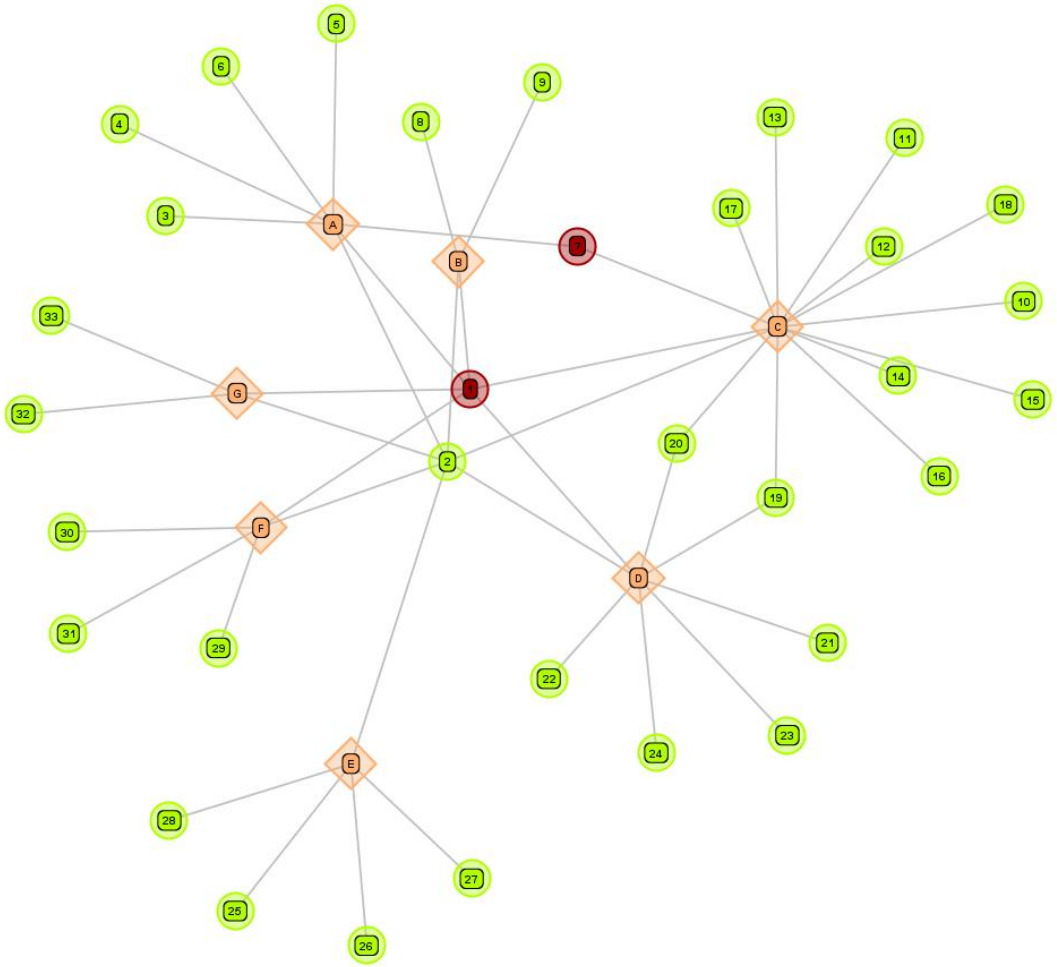


Top search results are the most relevant

KU LEUVEN

Gotcha! Propagation algorithm

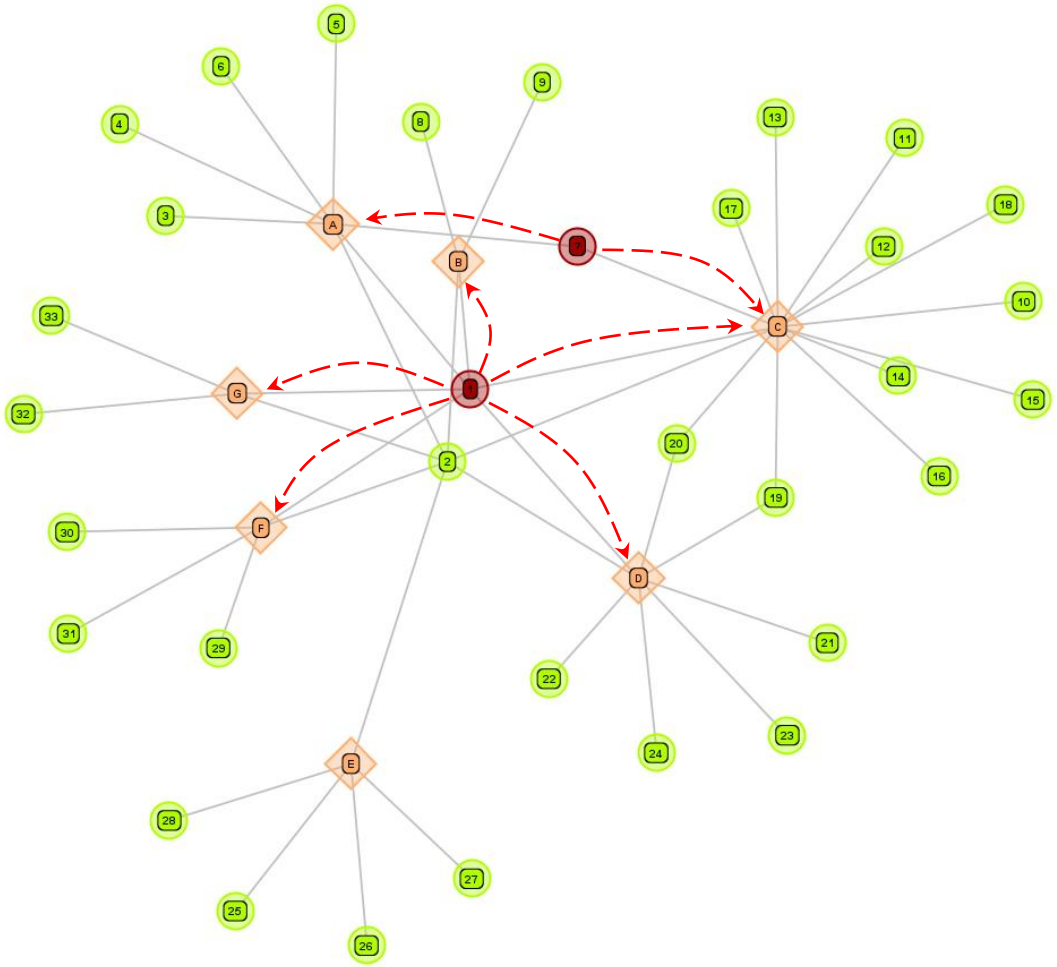
- Fraud propagation through the network: iterative procedure



KU LEUVEN

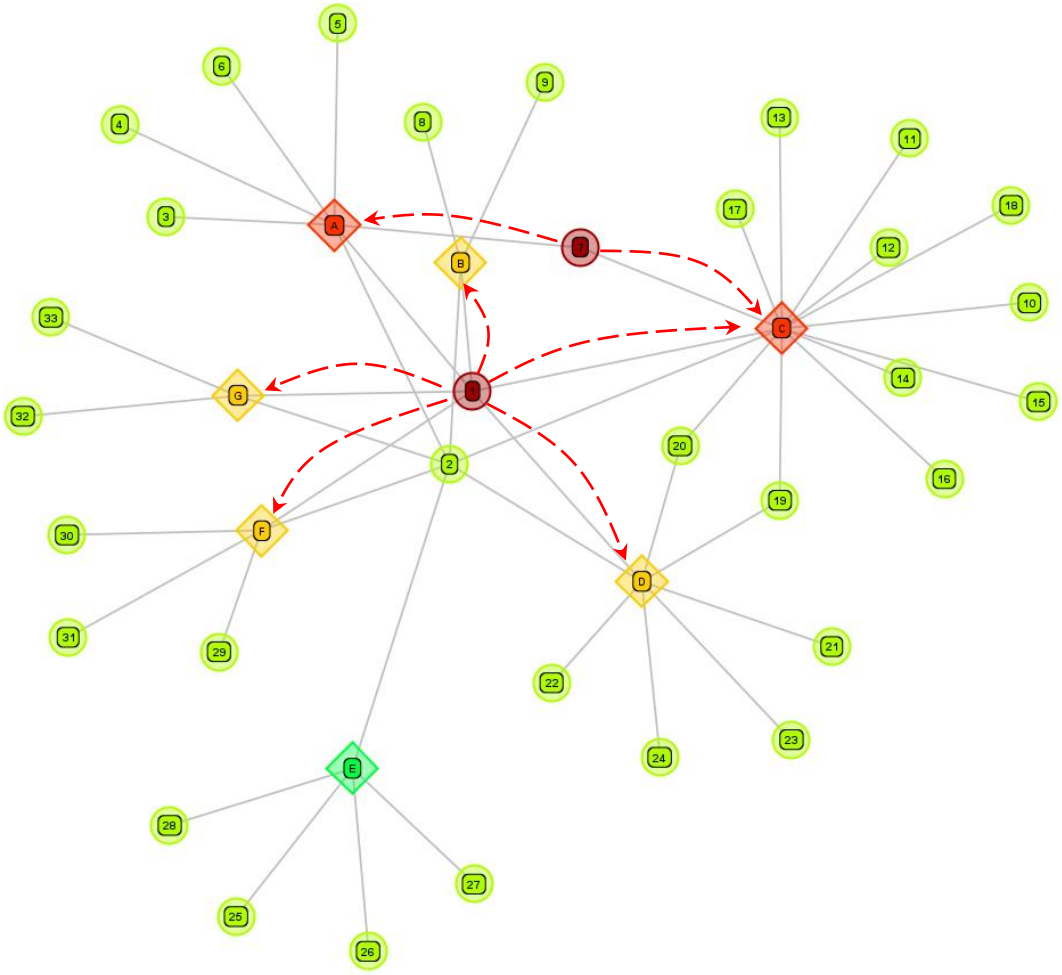
Gotcha! Propagation algorithm

- Fraud propagation through the network: iterative procedure



Gotcha! Propagation algorithm

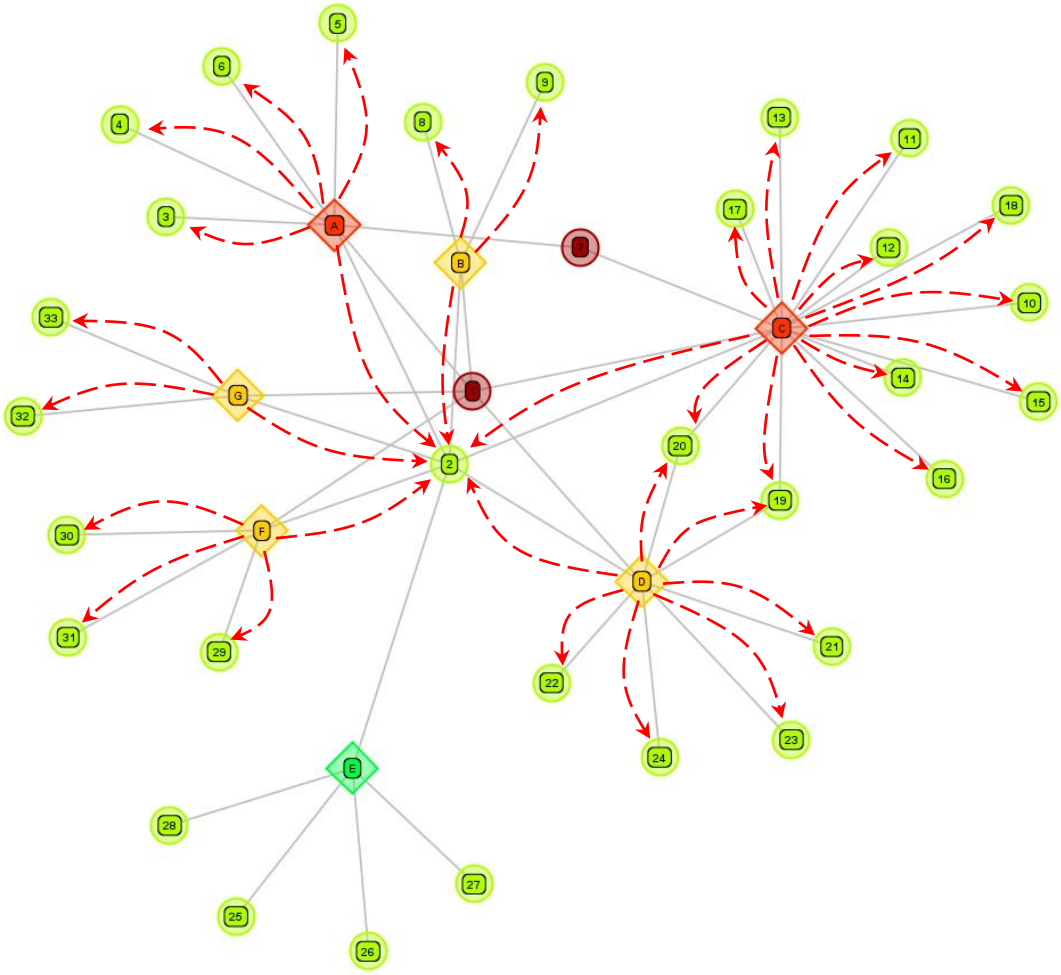
- Fraud propagation through the network: iterative procedure



KU LEUVEN

Gotcha! Propagation algorithm

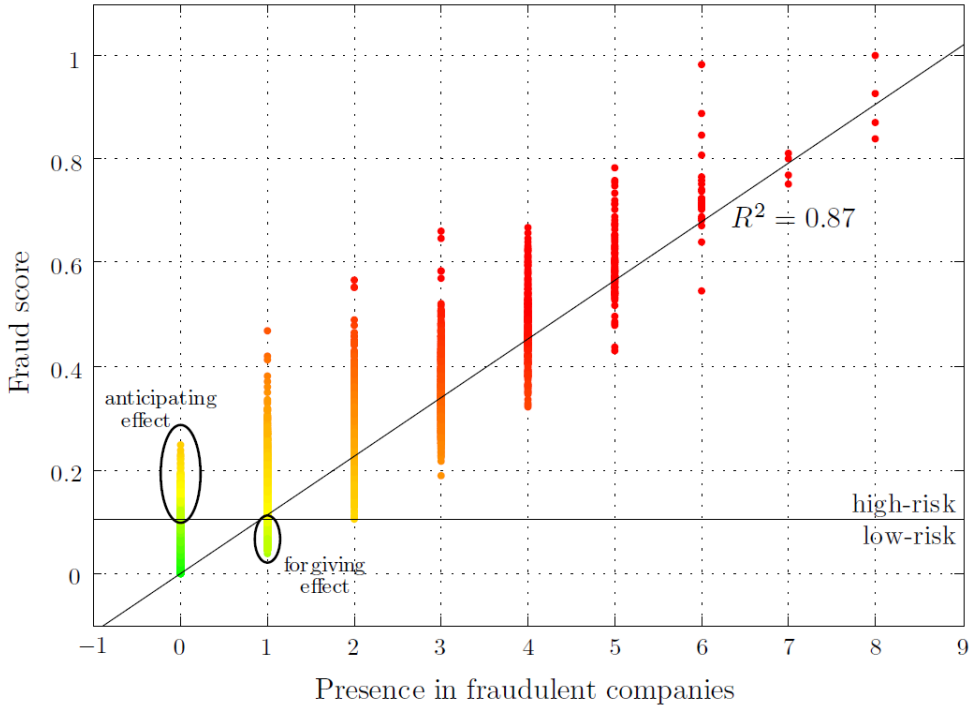
- Fraud propagation through the network: iterative procedure



Gotcha! Propagation algorithm

- Google's PageRank algorithm is extended for fraud detection

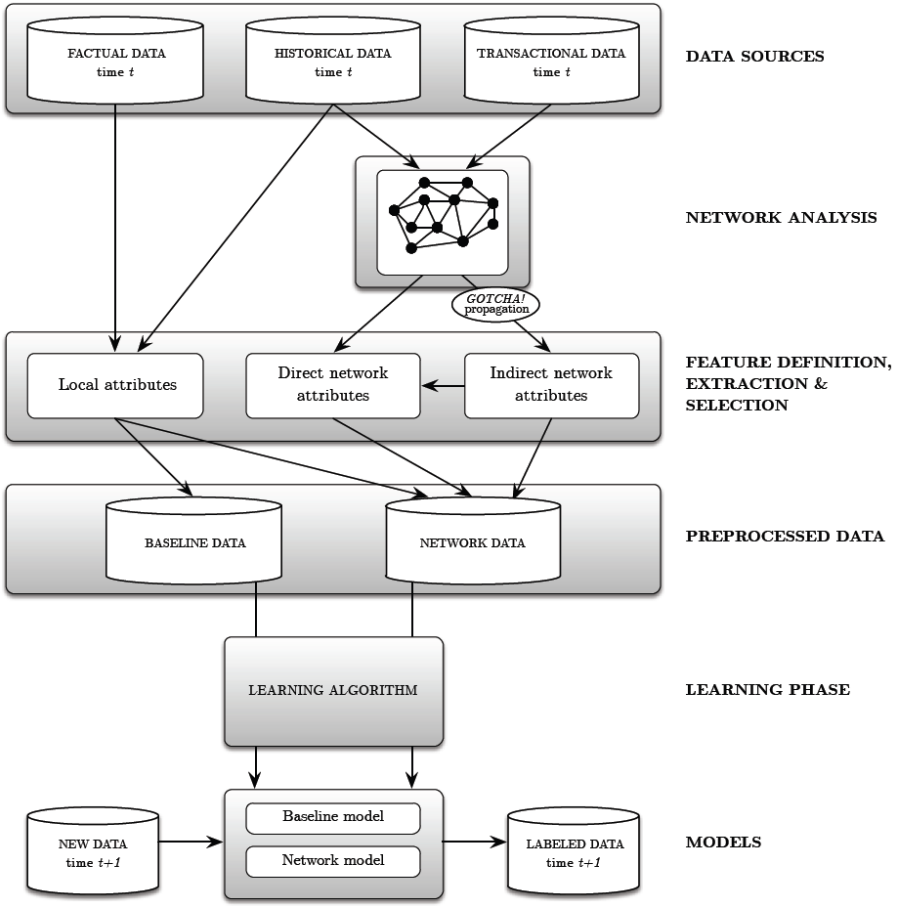
- Output:
  - Initial exposure score for legitimate companies / transactions
  - Riskiness of resources / merchants and credit card holders



KU LEUVEN

Gotcha! Outline

1. Data sources
2. Network analysis
3. Feature definition, extraction and selection
4. Learning phase
5. Detection and prevention model



KU LEUVEN

Gotcha! Featurization

- **Intrinsic features**
  - Local behavior as if the entity was treated in isolation

  - *Social security fraud (company level):*
    - Demographic data: sector, age, financial statements…
  - *Credit card transaction fraud (transaction level):*
    - RFM variables: does the transaction comply with normal customer behavior

- **Direct network features**
  - Egonet-based features (first-order neighborhood)

- **Indirect network features**
  - Gotcha! propagation algorithm: exposure scores

Gotcha! Featurization

- **Intrinsic features**
  - Local behavior as if the entity was treated in isolation

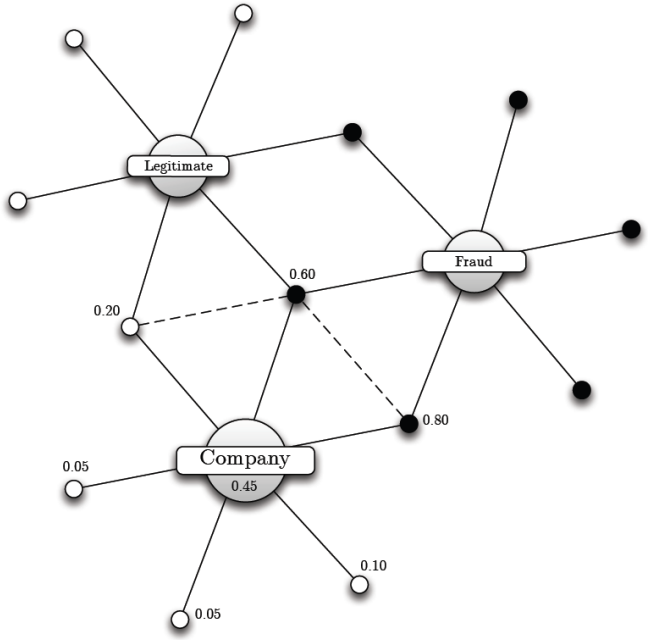
- **Direct network features**
  - Egonet-based features (first-order neighborhood)

  - *Social security fraud (company level):*
    - Company is linked to its resources: aggregate resource characteristics
  - *Credit card transaction fraud (transaction level):*
    - Transaction is linked to a merchant and credit card holder: include their exposure scores

- **Indirect network features**
  - Gotcha! propagation algorithm: exposure scores

Gotcha! Featurization

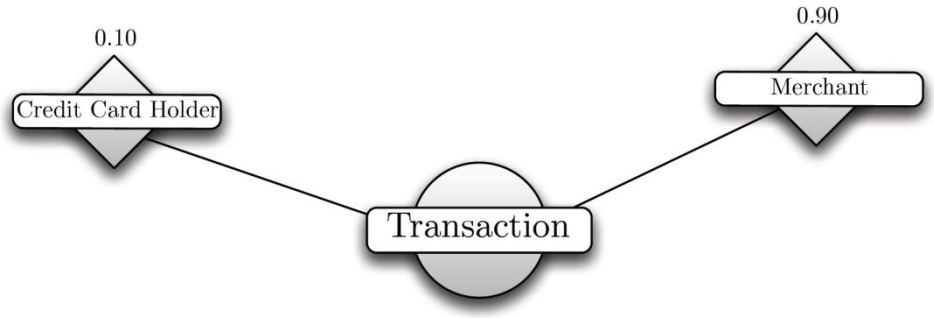
- **Direct network features**
  - Egonet-based features (first-order neighborhood)
  - *Social security fraud (company level):*
    - Company is linked to its resources: aggregate resource characteristics
    - *Triangles:* resources previously involved in the same (fraudulent) company



KU LEUVEN

Gotcha! Featurization

- **Direct network features**
  - Egonet-based features (first-order neighborhood)
  - *Credit card transaction fraud (transaction level):*
    - Transaction is linked to a merchant and credit card holder: include their exposure scores



KU LEUVEN

Gotcha! Featurization

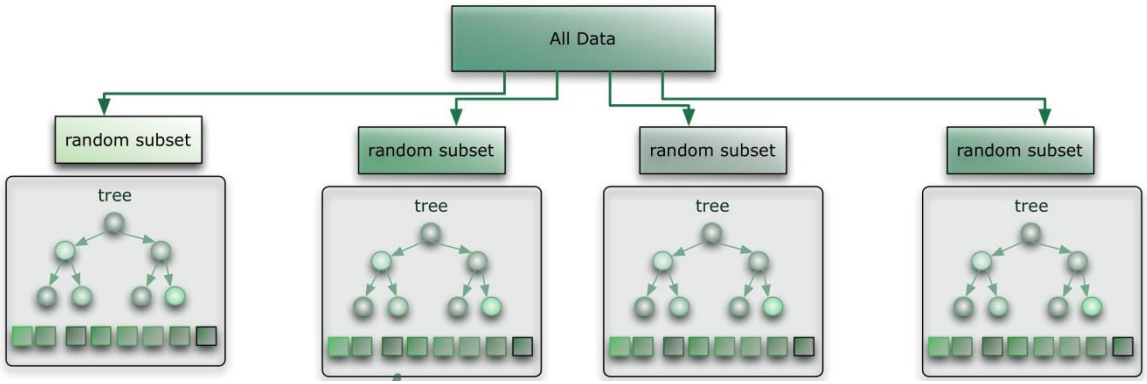
- **Intrinsic features**
  - Local behavior as if the entity was treated in isolation
- **Direct network features**
  - Egonet-based features (first-order neighborhood)
- **Indirect network features**
  - Gotcha! propagation algorithm: exposure scores
  - *Social security fraud (company level):*
    - Company exposure score
  - *Credit card transaction fraud (transaction level):*
    - Transaction exposure score

KU LEUVEN

Gotcha! Learning phase

- Ensemble learning:
  - Typically many features due to network analysis
  - Ensemble methods randomly select features and use a voting system for the final fraud probability

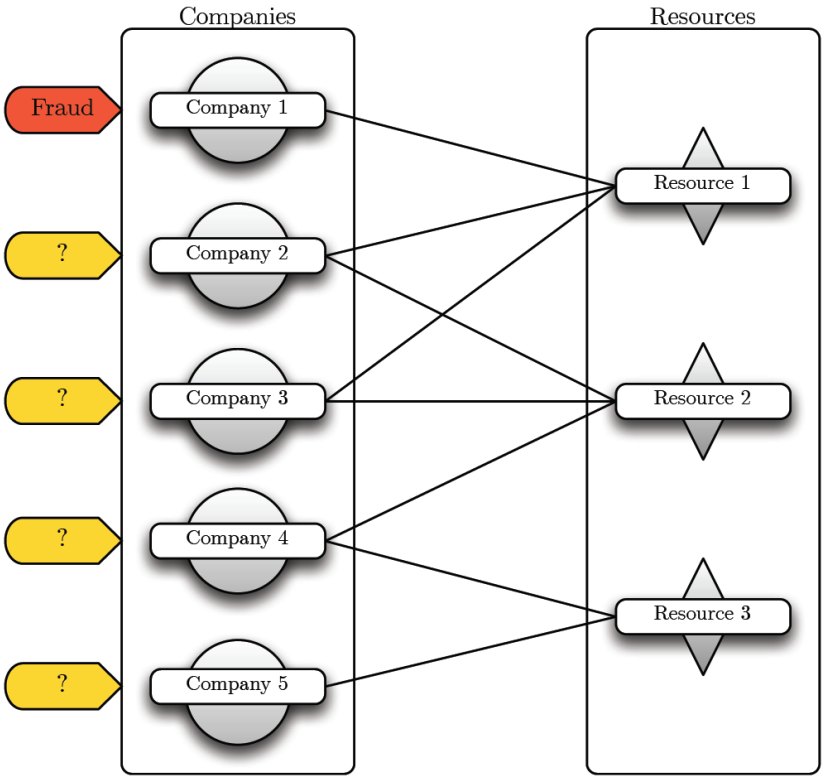
  - *Methods*:
  - Random forest
  - Logistic forest



KU LEUVEN

Detection model

- Social security fraud



KU LEUVEN

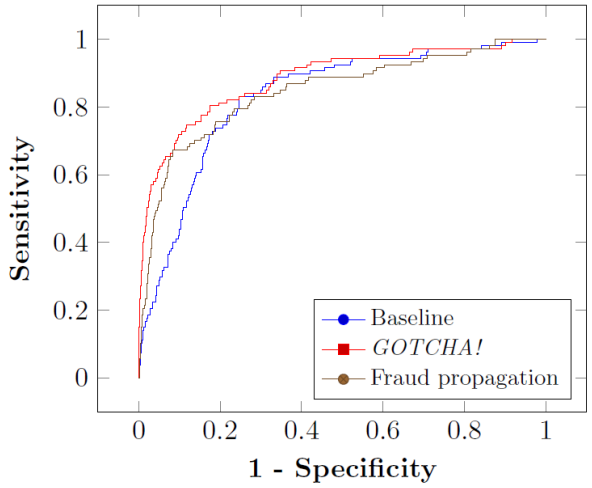
Detection model

- Social security fraud

		Total	ST Fraud	MT Fraud	LT Fraud	Fraud after analysis	Total Fraud	Bankrupt	Non-Active	Active	% detected
Year t_1	Baseline	100	10	5	2	7	**24%**	24	16	36	**48%**
	GOTCHA!	100	24	7	6	7	**44%**	27	5	24	**71%**
Year t_2	Baseline	100	6	2	10	1	**19%**	14	16	51	**33%**
	GOTCHA!	100	18	5	10	4	**37%**	24	7	32	**61%**
Year t_3	Baseline	100	11	1	1	0	**13%**	4	4	79	**17%**
	GOTCHA!	100	29	9	6	0	**44%**	15	4	37	**59%**

ROC curve:

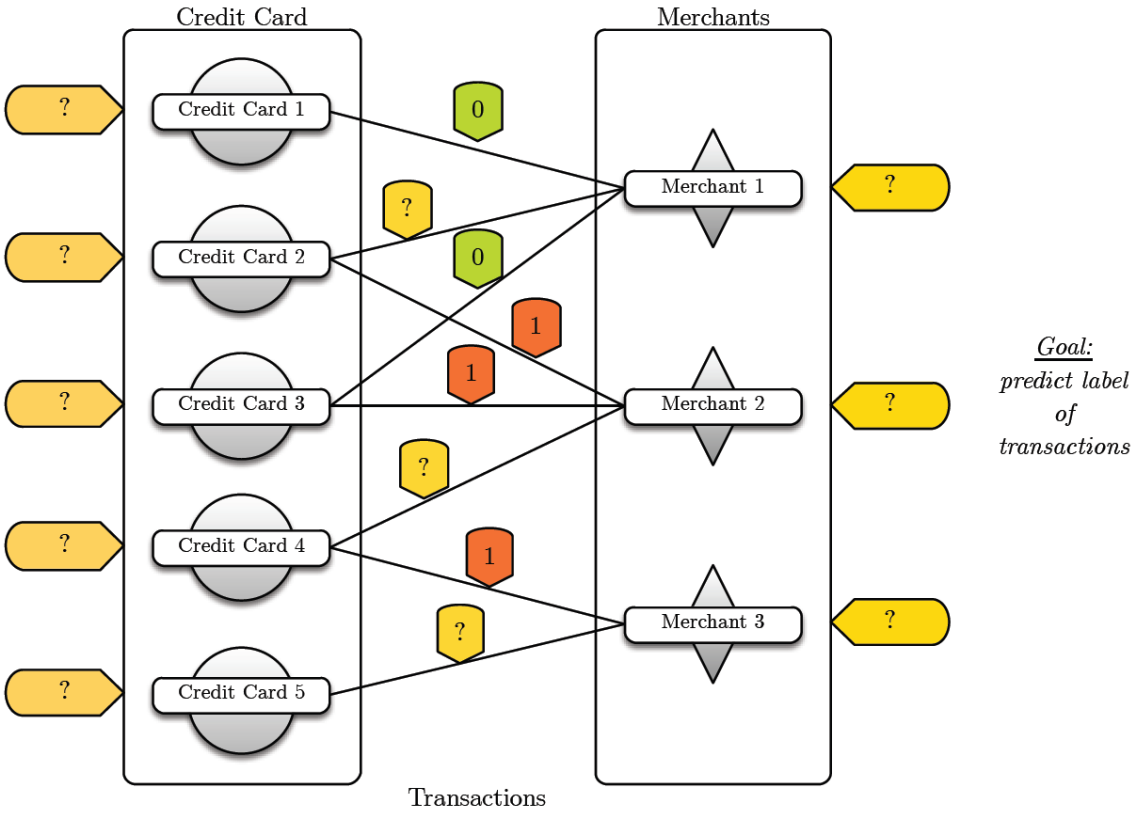
Out-of-time validation (Year t_2 - Year t_3)



KU LEUVEN

Detection model

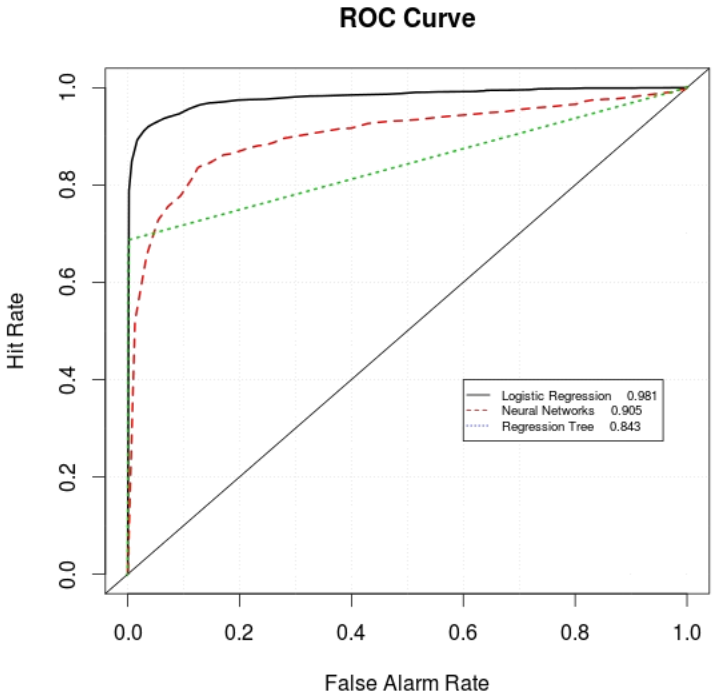
- Credit card fraud

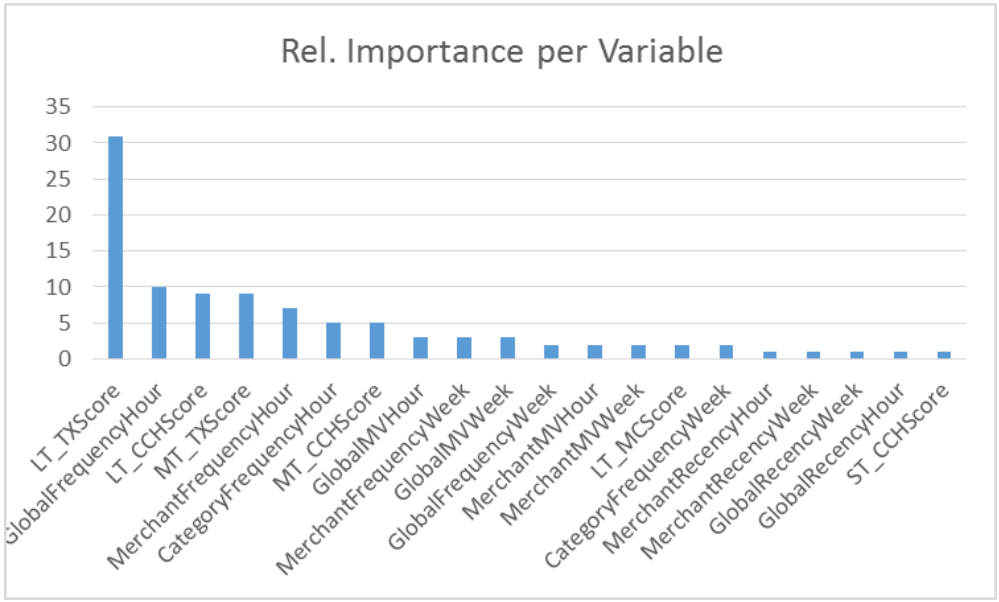


KU LEUVEN

Detection model

- Credit card fraud



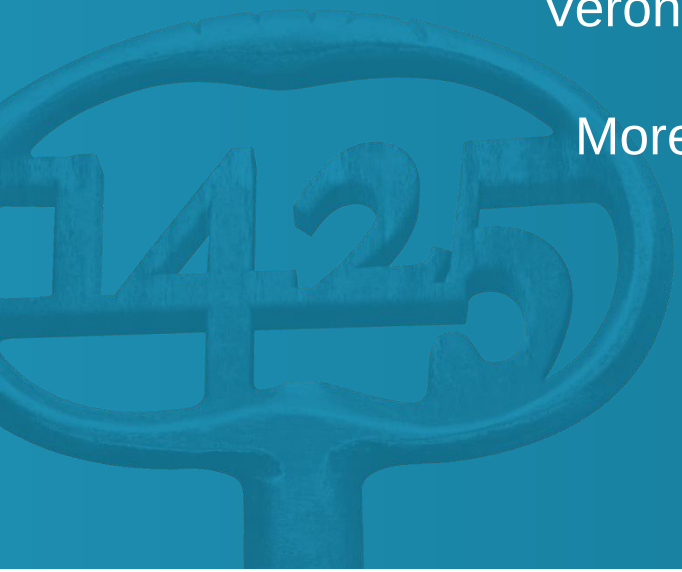


KU LEUVEN

Contact

Bart.Baesens@kuleuven.be
Veronique.VanVlasselaer@kuleuven.be

More info: www.dataminingapps.com
(mini lecture series)



KU LEUVEN

2. The best way to create high-performing fraud detection models is to use:

A. a combination of both social network & intrinsic variables
B. social network variables
C. intrinsic variables

Tweet your answer:

Example: @spicyanalytics 2C

Start of your tweet | Question # | Your answer

Prizes to win:

1st prize: a ticket for Analytics 2015
2nd prize: a book of Prof Bart Baesens: "Analytics in a big data world"
3rd to 30th prize: chocolates with pepper

Winners will be contacted post-Forum !