Network Analysis of the *Antifraud Integrated Archive* for Fraud Detection

Andrea Consiglio[1]

joint work with

Michele Tumminello[1], Riccardo Cesari[2], Fabio Farabullini[2], Pietro Vassallo[3]

[1]Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli studi di Palermo

[2]Istituto per la Vigilanza sulle Assicurazioni (IVASS)

[3]Bank of Italy

Summary

- The Antifraud Integrated Archive (AIA)
- Bipartite Networks and statistically validated networks
- Network indicators
- Conclusions

Paper

- Tumminello, M., Consiglio, A., Vassallo P., Cesari, R. and Farabullini, F. *"Insurance fraud detection: A statistically validated network approach",* Journal of Risk and Insurance, 2023.

Anti-fraud **I**ntegrated **A**rchive (AIA)

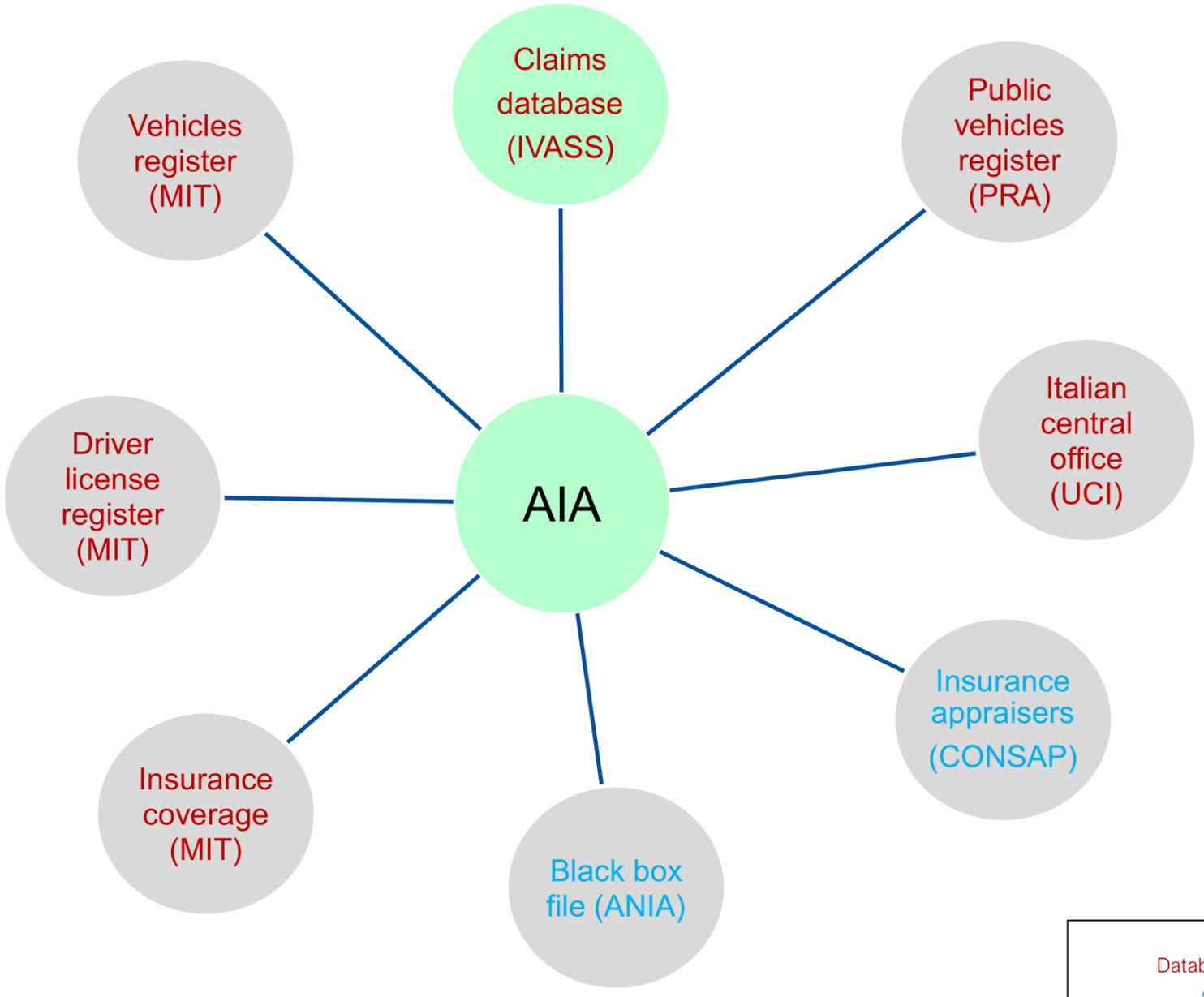
In **2012** and **2017 laws** passed, which introduce relevant innovations for fighting frauds

In particular, such laws allowed **IVASS** to collect information from **external databases** in order to increase the information available for **anti-fraud activity**

Consequently, IVASS has implemented a new tool called

ANTI-FRAUD INTEGRATED ARCHIVE (AIA)

AIA: stage 1


Claims database (IVASS)
Vehicles register (MIT)
Public vehicles register (PRA)
Italian central office (UCI)
Driver license register (MIT)
AIA
Insurance appraisers (CONSAP)
Insurance coverage (MIT)
Black box file (ANIA)
Database foreseen by the law
Additional database

Indicators and scores (before network tools)

- Binary Indicators (on/off)
- Built on the bases of **recurrences** and **cross-check** criteria
- Different weight according to the relevance in anti-fraud activity

Indicators and scores (before network tools)

Score vehicle

Score people **directly** involved

Score people **indirectly** involved

Score others aspects

Overall Score

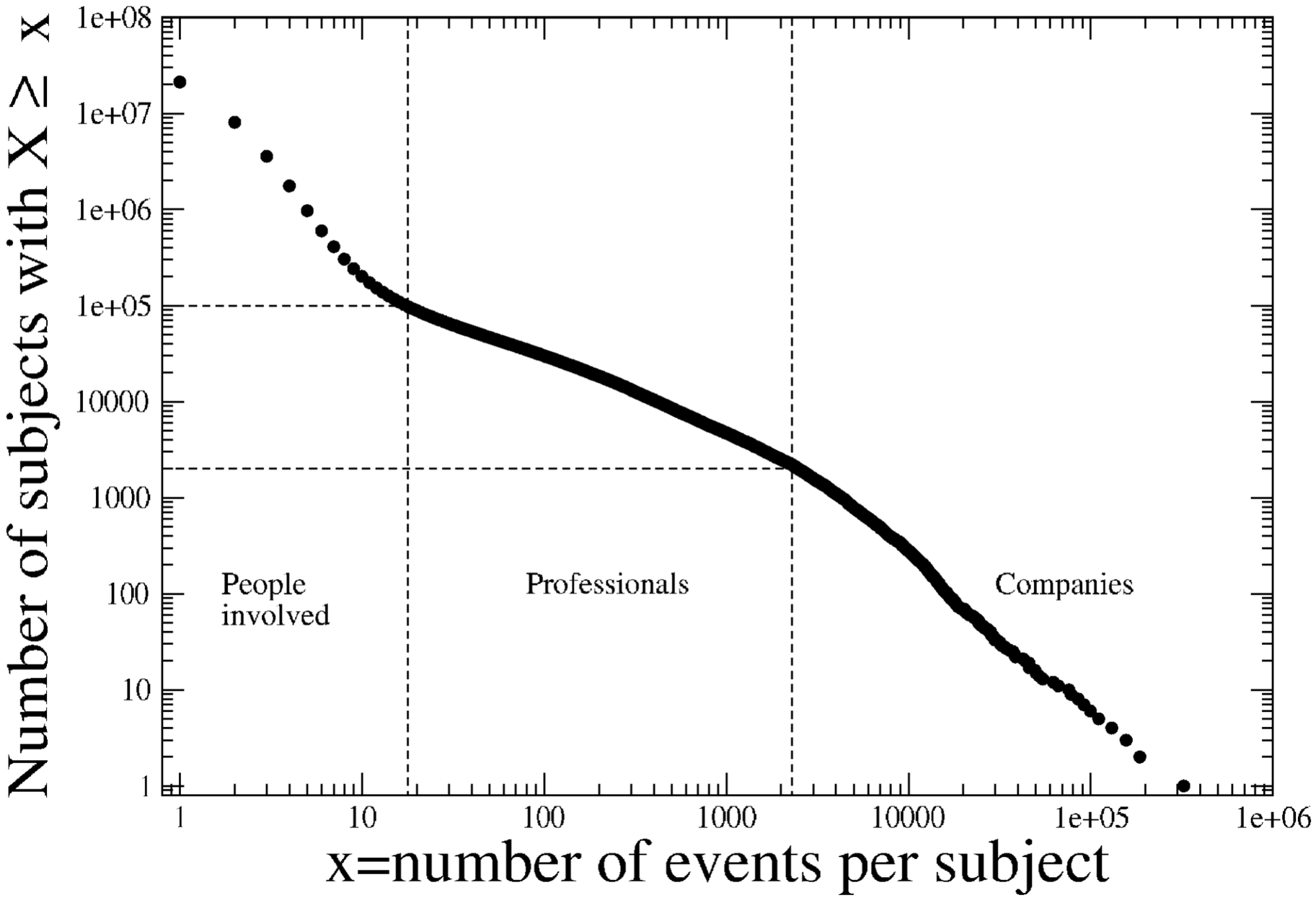
AIA: stage 2

Network Analysis

Big Data: AIA

- Time period: 2011-2018
- About 18.5 million car accidents
- About 23.5 million individuals and companies
- About 19 million vehicles
- A *Data Lake* ….

Heterogeneity of subjects

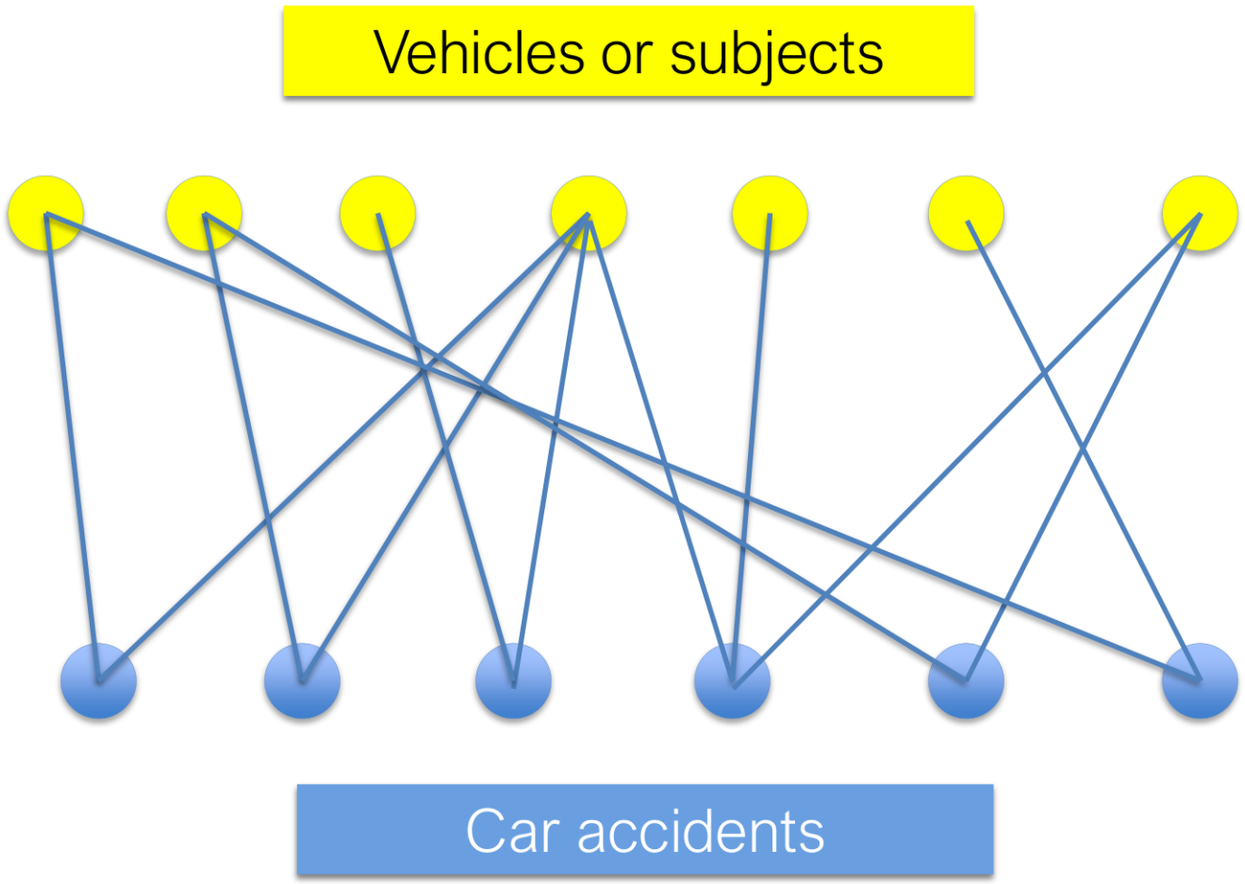


Objectives

- Uncover patterns in the data that suggest fraudulent activity.

- Identify organized groups of perpetrators.

Bipartite networks

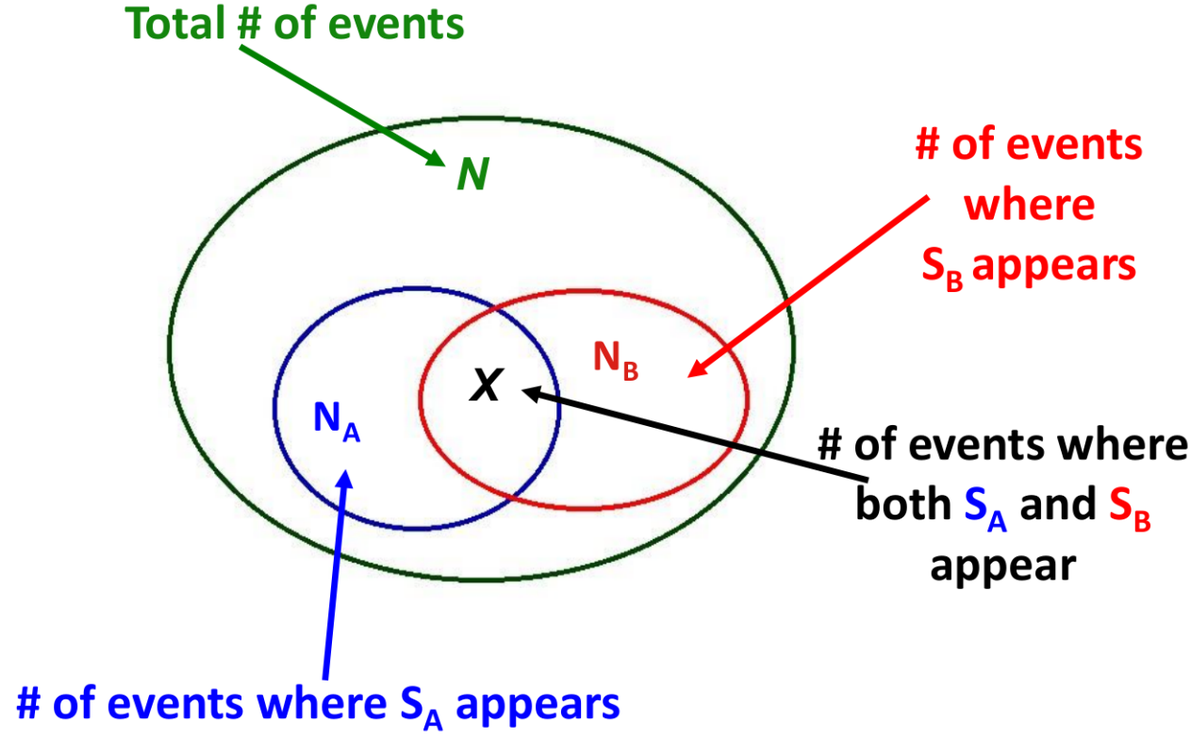


Null hypothesis

One does not choose the counterpart in an accident

A statistical validation of co-occurrence

Suppose there are **N** events in the investigated set. We want to statistically validate the co-occurrence of subject S_A and subject S_B in **X** events against a null hypothesis of random co-occurrence. Suppose that the number of events where S_A (S_B) appears is N_A (N_B), whereas the number of events where both S_A and S_B appear is ***X***.



The question that characterizes the null hypothesis is:
what is the probability that number X occurs by chance?

Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN (2011) Statistically Validated Networks in Bipartite Complex Systems. PLOS ONE 6(3): e17994. doi:10.1371/journal.pone.0017994
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017994

Hypergeometric distribution and Statistically Validated Networks

p-value associated with a detection of co-occurrences ≥ X:

$$p = \sum_{i=X}^{\min(N_A,N_B)} \frac{\binom{N_A}{i}\binom{N-N_A}{N_B-i}}{\binom{N}{N_B}}$$

- Count the total number of tests: *T*

- Arrange ***p-values*** in increasing order.

- Set a link between two vertices if the associated p-value satisfies one of the following inequalities

Bonferroni correction : $p - value_{(k)} < \frac{\alpha}{T}$ → **Bonferroni Network**

Holm-Bonferroni correction : $p - value_{(k)} < \frac{\alpha}{T-k}$ → **Holm-Bonferroni Network**

FDR correction : $p - value_{(k)} < \frac{\alpha\, k}{T}$ → **FDR Network**

Type I error control: false positive links

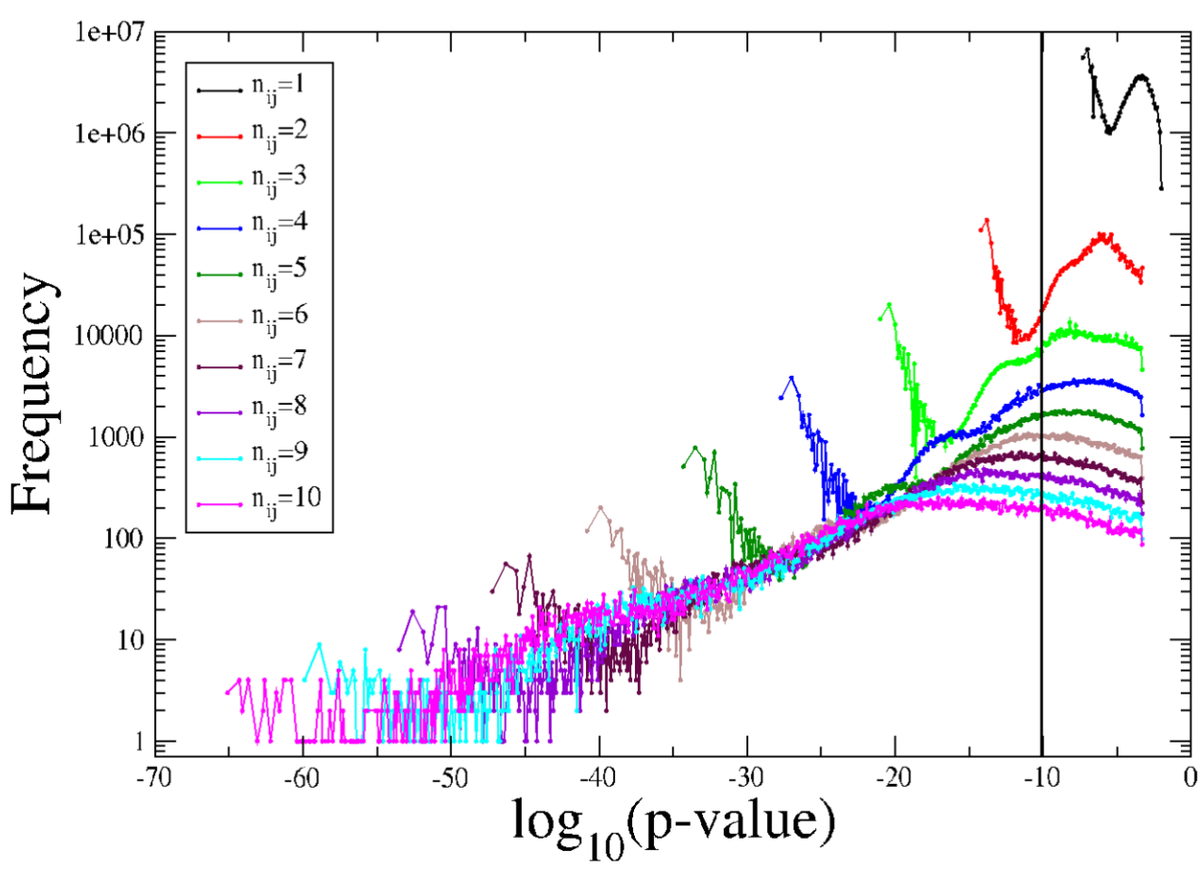
Proposition: the probability that a false positive link is set in the **Bonferroni network** is smaller than α.

Co-occurrences might be dependent

Bonferroni network

- It's the most conservative statistically validated network
- The threshold is independent of p-values
- A **co-occurence** equal to **1** is not statistically significant, provided that the number of links, E, in the co-occurrence network is larger than the number of nodes, N, in the projected set, times α

$$p - value(n_{AB} = 1, N_A, N_B, N) \geq p - value(n_{AB} = 1, 1, 1, N) = \frac{1}{N} > \frac{\alpha}{E}$$



Distinguishing between subjects and vehicles

	Nodes	Links	Connected components (CC)	Size of largest CC
Bonferroni network of **subjects***	2,016,505	1,919,897	638,878	**651,267**
Bonferroni network of **vehicles***	112,771	61,311	54,563	**12**

*Subjects and vehicles recorded in the white list have been excluded from the analysis

Bonferroni network of subjects: largest communities

Community ID	Years over-expressed	Regions over-expressed	Provinces over-expressed
1	2015, 2016	SARDEGNA, LOMBARDIA, LAZIO	VA, TV, TP, TO, SS, RM, RN, RG, PO, PT, PE, PV, PD, MI, LO, LC, LT, CO, CL, CA, BG, MB, OG, VI, VR, AG
2	2011, 2012	CAMPANIA*, NA	NULL, SA, AV, NA, CE
3	-	TOSCANA*, NA	NULL, SI, PO, PT, PI, AR, LU, FI
4	-	PIEMONTE*, VALLE_D'AOSTA	VC, TO, AT, AO, CN, BI
5	-	BASILICATA, PUGLIA*, NA	NULL, BA, TA, PZ, MT, FG, BR, BT
6	-	FRIULI_VENEZIA_GIULIA, VENETO*	VE, UD, TV, RO, PN, PD, FE, VI, VR, BL
7	-	SICILIA*	TP, PA, AG
8	-	LAZIO*	RM, RI, LT, VT
9	-	SICILIA*, NA	NULL, SR, RG, ME, EN, CT, CL
10	-	EMILIA_ROMAGNA*	RN, RA, OR, MO, FC, FE, BO
11	2015, 2016	LAZIO*	RM, RI, LT, FR, VT
12	2011	FRIULI_VENEZIA_GIULIA, VENETO	VE, UD, TV, PN, PD, NO, GO, VI, BL
13	-	LIGURIA, NA	NULL, SV, SP, IM, GE, AL
14	-	LAZIO, NA	NULL, RM, LT, VT
15	2015	CAMPANIA*	SA, AV, NA, CE
17	-	EMILIA_ROMAGNA*, NA	NULL, RE, PR, MO, MN, FE, BO
23	2016	LOMBARDIA	VA, PV, MI, LO, LC, CR, CO, BG, MB
25	-	LOMBARDIA, NA	PC, MN, LO, CR, BS, BG, VR

Are links robust to time-space localization?

*Homogeneity larger than 90%

An indicator of link-robustness to localization

T=total number of events in the dataset (**T**=13,533,500 in AIA 10/2016)

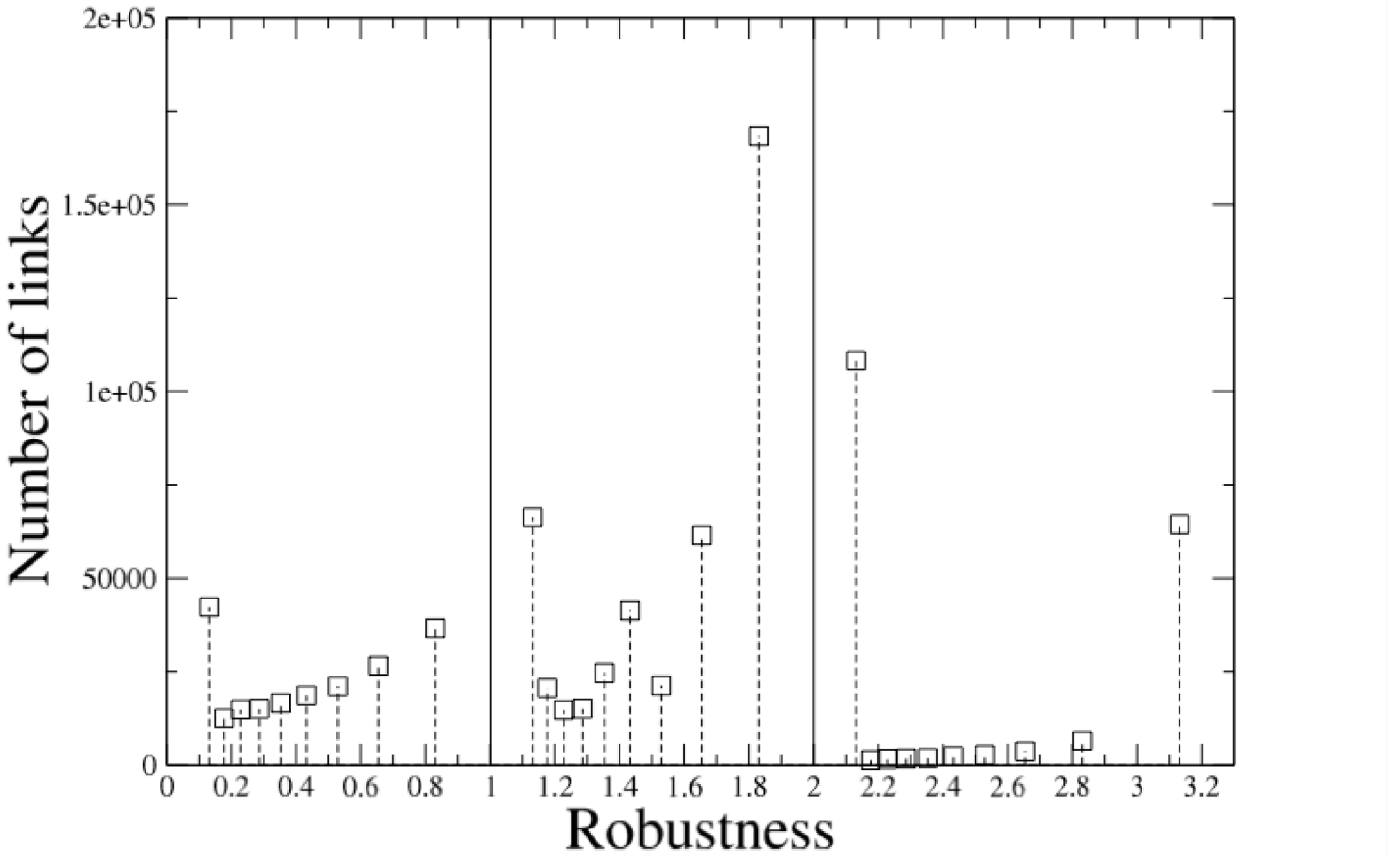
B=bonferroni threshold in the dataset (**B**=1.356e-10 in AIA 10/2016)

M(i,j)=Min(Q) such that p-value(n(i),n(j),n(i,j),Q)<**B**

Robustness indicator

$$R(i,j)=log_{10}(T)-log_{10}(M)$$

Bonferroni network: distribution of link-robustness (R>0.1)



Node (event, subject, vehicle) indicators of centrality

- Node degree
- Node total strength
- Node average strength
- Node betweenness

Mixed Event-subject indicators

Statistically Validated Bipartite Network

Construction: given the SVN of subjects (or vehicles), a bipartite network is reconstructed by

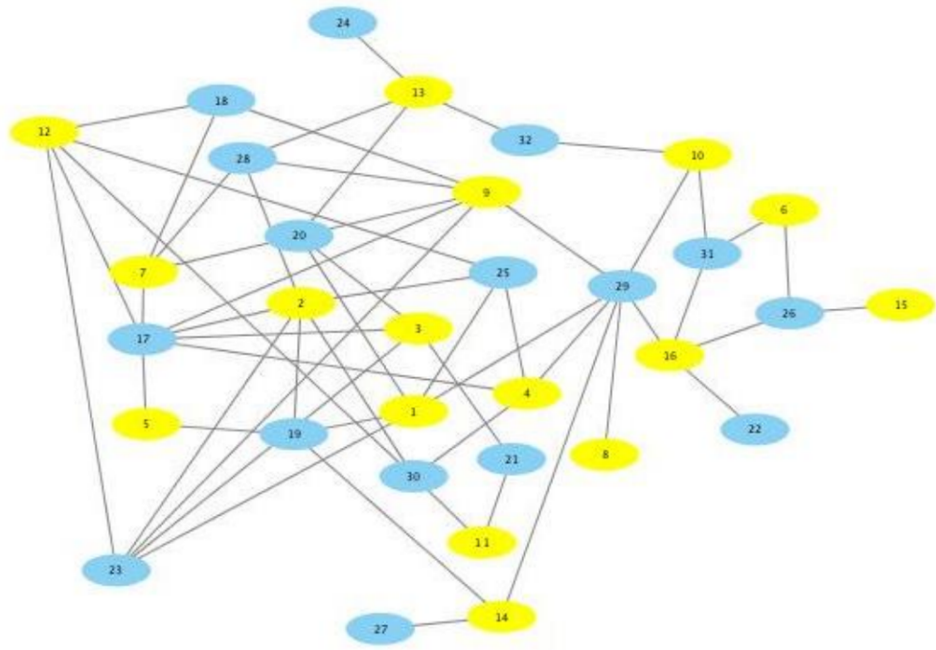
- selecting from the original bipartite network all of the ***event(i)-subject(j)*** pairs such that ***event(i)*** contributed to a **link in the SVN between *subject(j)*** and (at least) another subject.

- adding afterwards all of the subjects directly involved in the selected events.

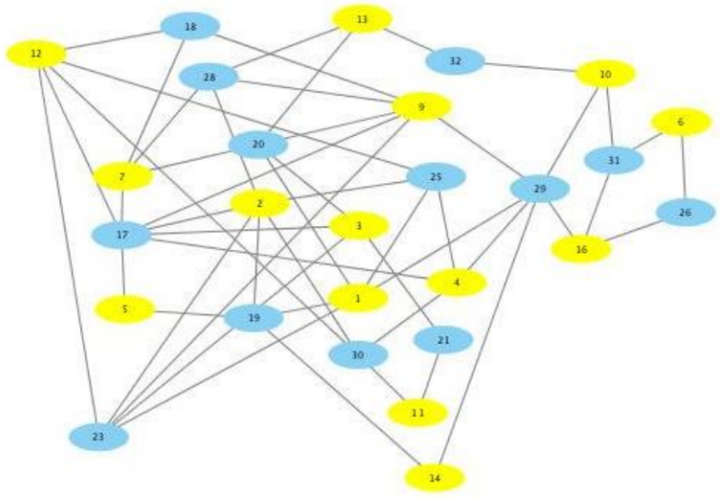
K-H core of a bipartite network

The K-H core of a bipartite network is the largest bipartite **subnetwork** such that nodes of Set A have degree at least K and nodes of set B have degree at least H.

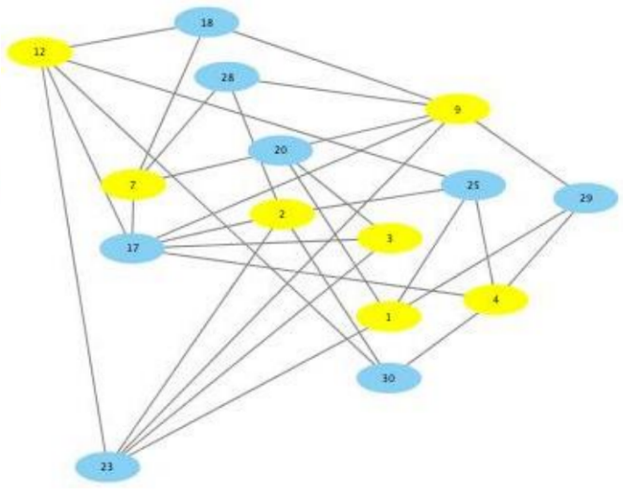
Bipartite network of Kids(blue)-toys(yellow)



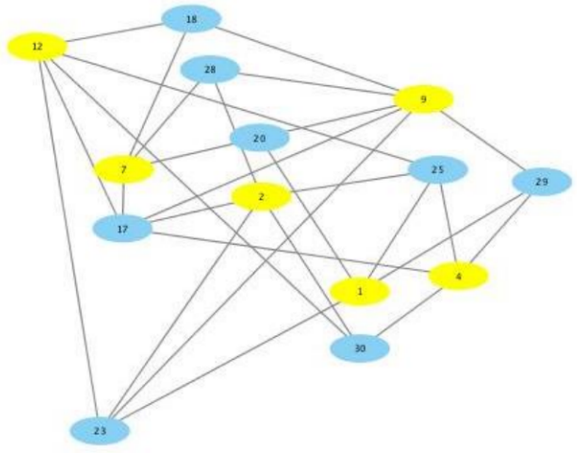
2-2 core



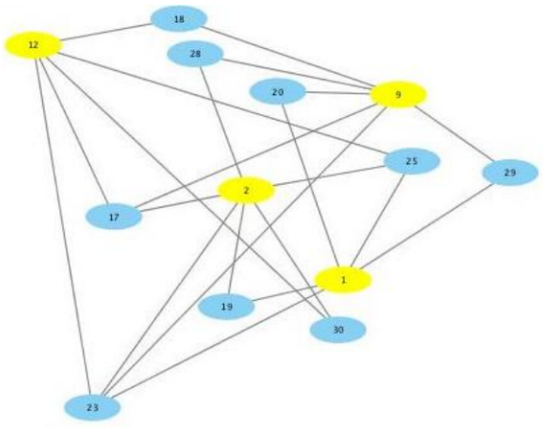
3-3 core



4-3 core



5-2 core



Network indicators: Mixed event-subject indicators of centrality: the **K-H core**

- Event oriented event-subject indicator:

$$KH_e(e,s) = \max(K) \text{ such that } (e,s) \in K-H \text{ core}$$

- Subject oriented event-subject indicator:

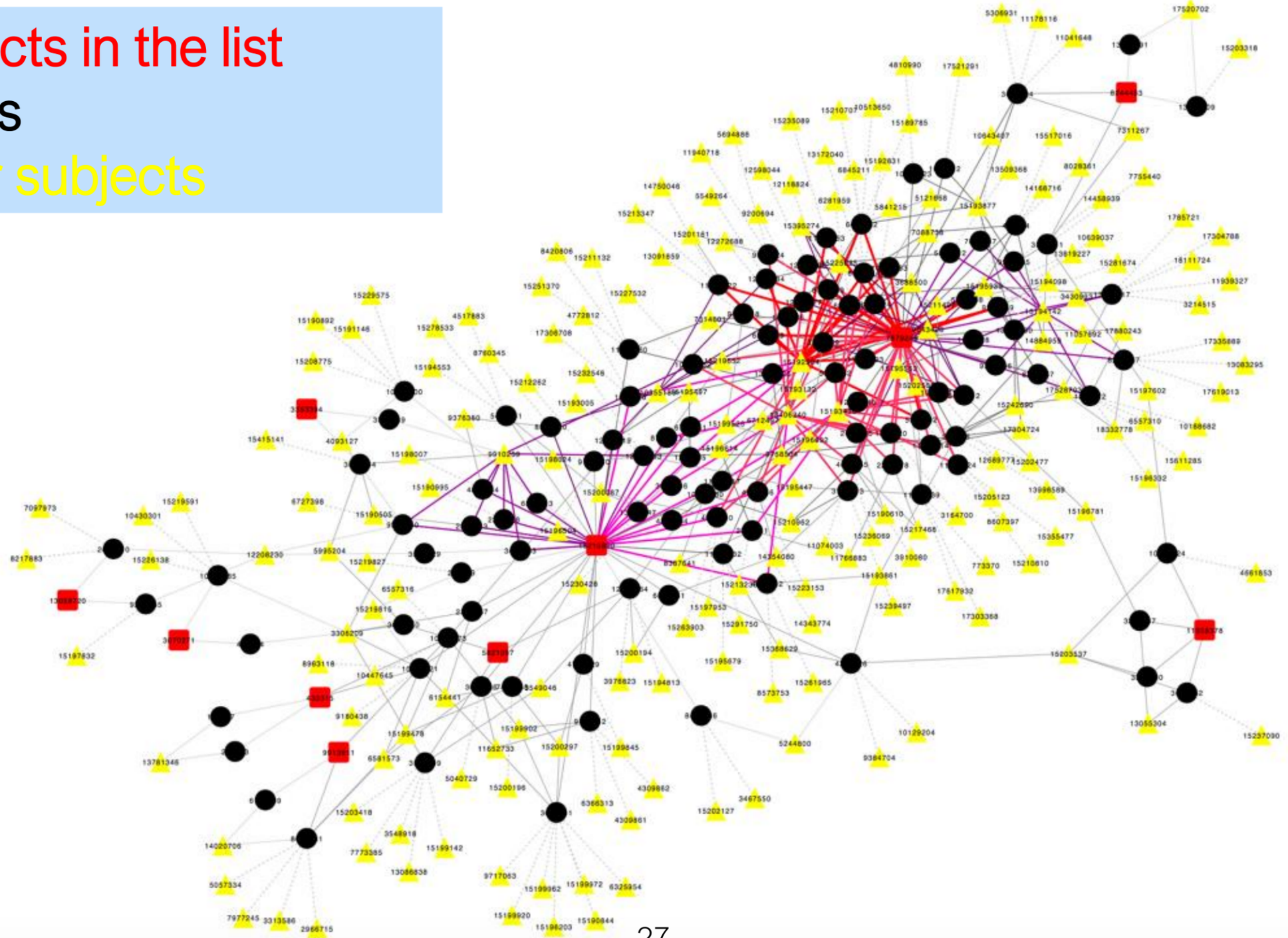
$$KH_s(e,s) = \max(H) \text{ such that } (e,s) \in K-H \text{ core}$$

- Balanced event-subject indicator:

$$KH(e,s) = \max(\sqrt{K \cdot H}) \text{ such that } (e,s) \in K-H \text{ core}$$

Case study 1: a reported list of subjects

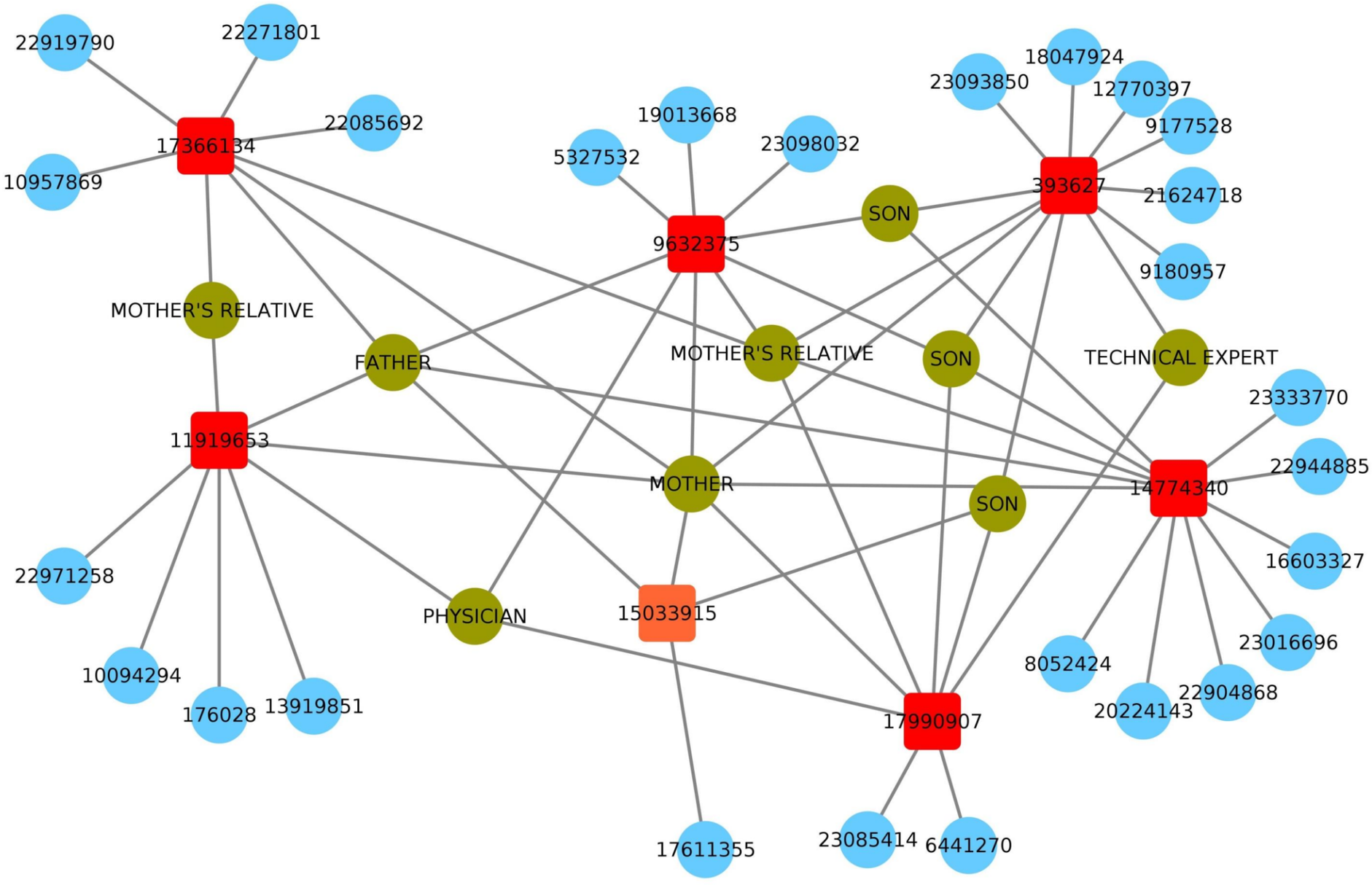
- Subjects in the list
- Events
- Other subjects



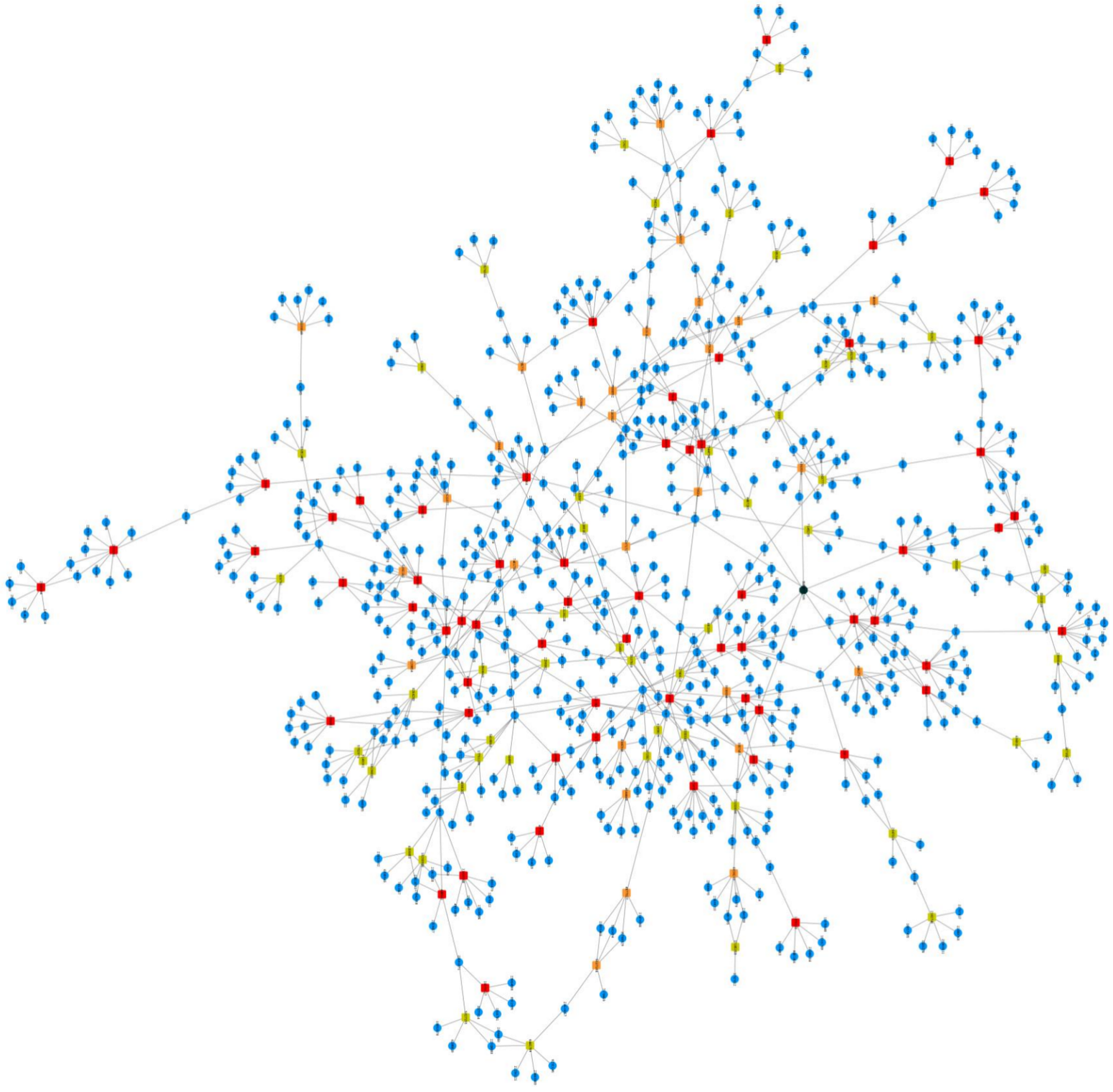
K-H shell: k-h(4,3)

H=3 events	**K=4 subjects**	Role
Event 1 Event 2 Event 3	Subject 1	Doctor in the list
Event 1 Event 2 Event 3	Subject 2	Doctor appointed by the company
Event 1 Event 2 Event 3	Subject 3	Medical consulting company
Event 1 Event 2 Event 3	Subject 4	Lawyer appointed by the counterpart

Case study 2: a community of relatives



Case study 3: legal identity theft



Subjects in the extended bipartite SVN

Legal subject

Accidents (anomaly: high)

Accidents (anomaly: intermediate)

Accidents (anomaly: low)

An integrated indicator

Many indicators, related to both the network (system) and the event
$\Rightarrow$ correlation is observed

Find new variables that are linearly independent

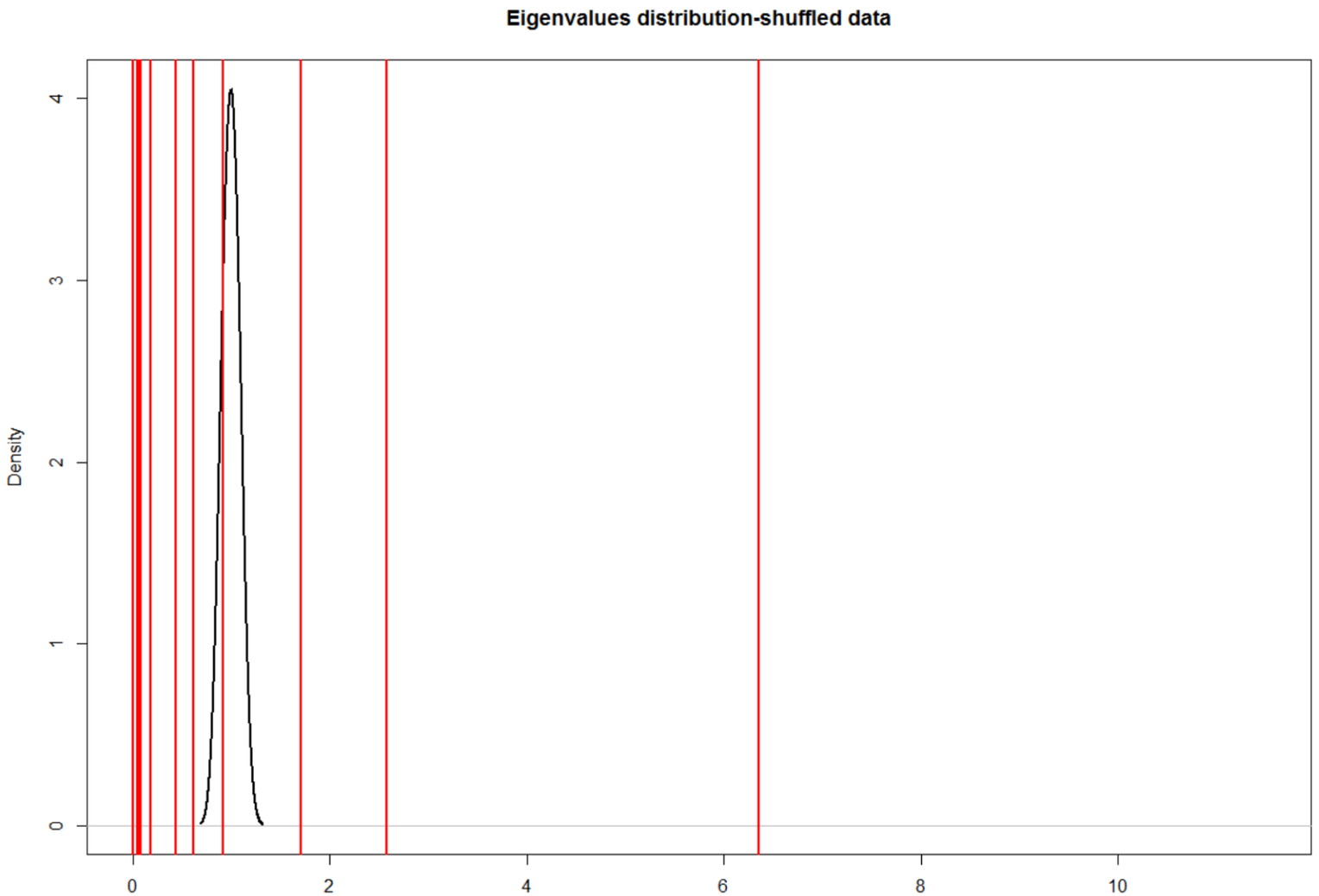
Select the "most informative" (RMT)

Integrated Indicator: modelling the selected composite variables

An integrated indicator: PCA & RMT

Cumulative percentage of variance explained by the PCs

1ª	2ª	3ª	4ª	5ª
51.6%	69.2%	81.1%	88.2%	92.0%



An integrated indicator: logit model

sample containing 6.753 events occured in Italy from 2014 to 2017

3.383 events randomly sampled from AIA

3.370 reported events

Asymmetric approach, cause-effect

What is the classification ability of the principal components?

Estimation of a logit model to estimate coefficients

$$logit\{\pi\} = \alpha_1 * CP_1 + \alpha_2 * CP_2 + \alpha_3 * CP_3 + \alpha_4 * CP_4$$

with π the probability of belonging to reported events

An integrated indicator: logit model

sample containing 6.753 events occured in Italy from 2014 to 2017

3.383 events randomly sampled from AIA

3.370 reported events

Out of sample validation

Initial dataset partitioned in two parts.

80% (5402 units) forms the training set.

20% (1351 units) the test set.

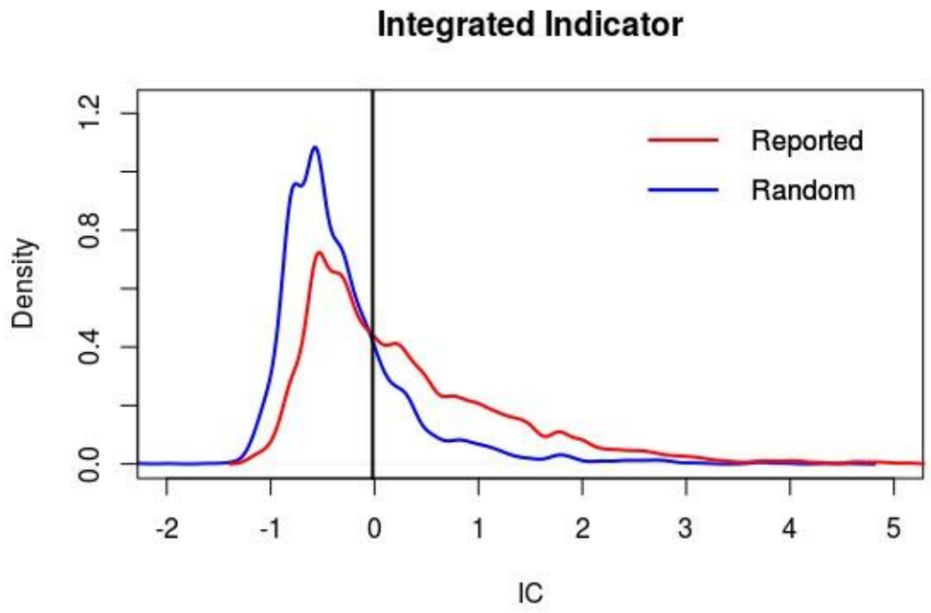
An integrated indicator: selection of the threshold

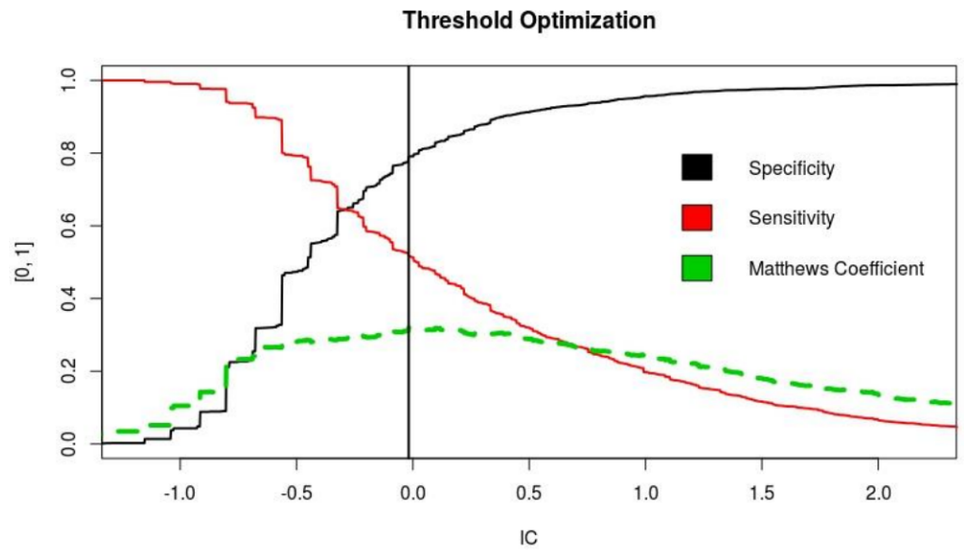
Initial dataset partitioned in two parts.

80% (5402 units) forms the training set.

20% (1351 units) the test set.

Maximize the Matthews Correlation Coefficient in the training set to select the threshold x_0



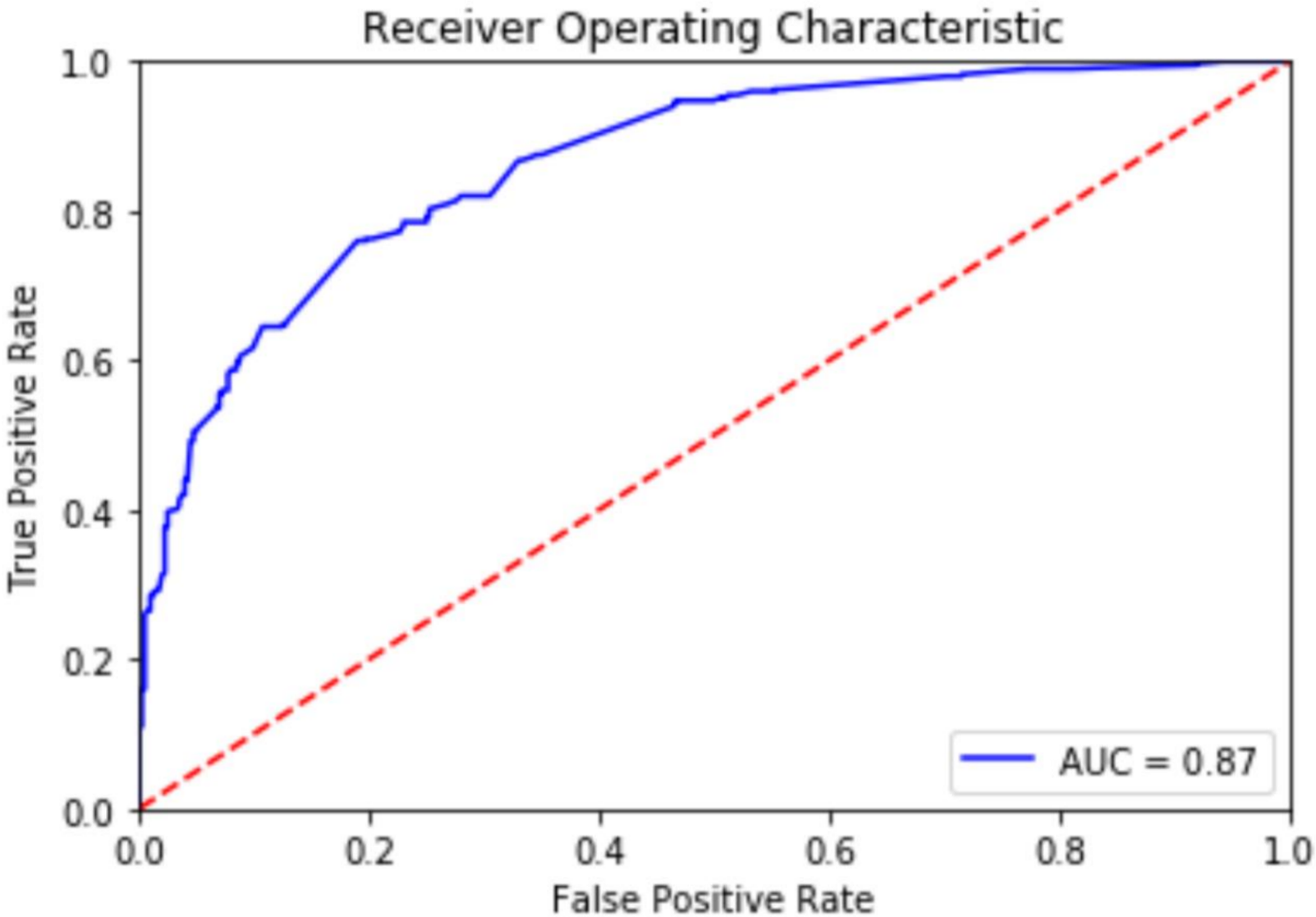


An integrated indicator: logit model

400 frauds VS a stratified random sample of 400 accidents

Out of sample Validation

AUC=0.87



An integrated indicator: performance

Initial dataset partitioned in two parts.

80% (5402 units) forms the training set.

20% (1351 units) the test set.

Maximize the Matthews Correlation Coefficient in the training set to select the threshold x_0=-0.018

Results (out of sample):

	Random	Reported
$X \leq x_0$	82% (3%)	47% (3%)
$X > x_0$	18% (3%)	53% (3%)
	100%	100%

AIA: stage 3

- Three node motifs
- Network analysis for data quality

Motifs: the heuristics

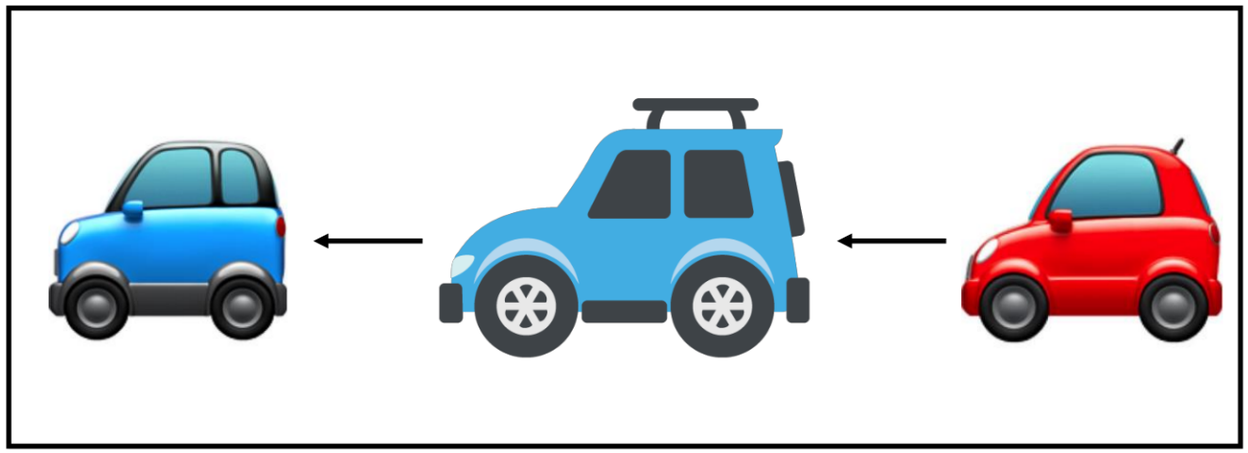
- Criminal specialization
- Some types of crime require cooperation
- Cooperating with a criminal intent requires secrecy and trust

Motifs

M Tumminello, C Edling, F Liljeros, RN Mantegna, J Sarnecki (2013) The Phenomenology of Specialization of Criminal Suspects. PLoS ONE 8(5): e64703. doi:10.1371/journal.pone.0064703

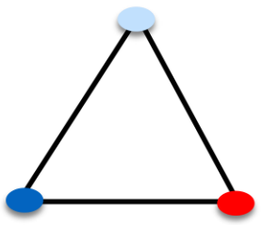
Motifs and anti-fraud

Not suspicious

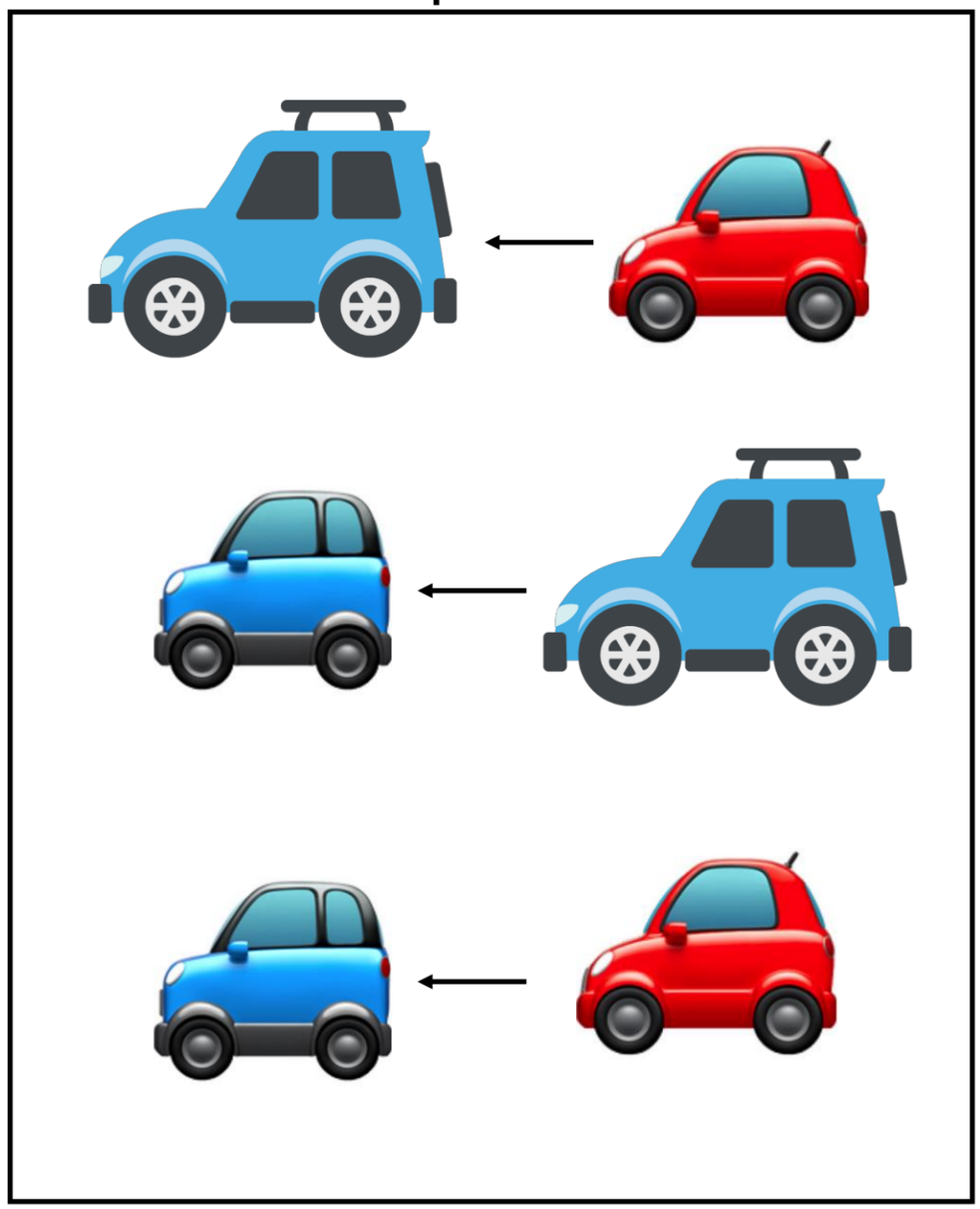


A single event involving three cars

Same projection

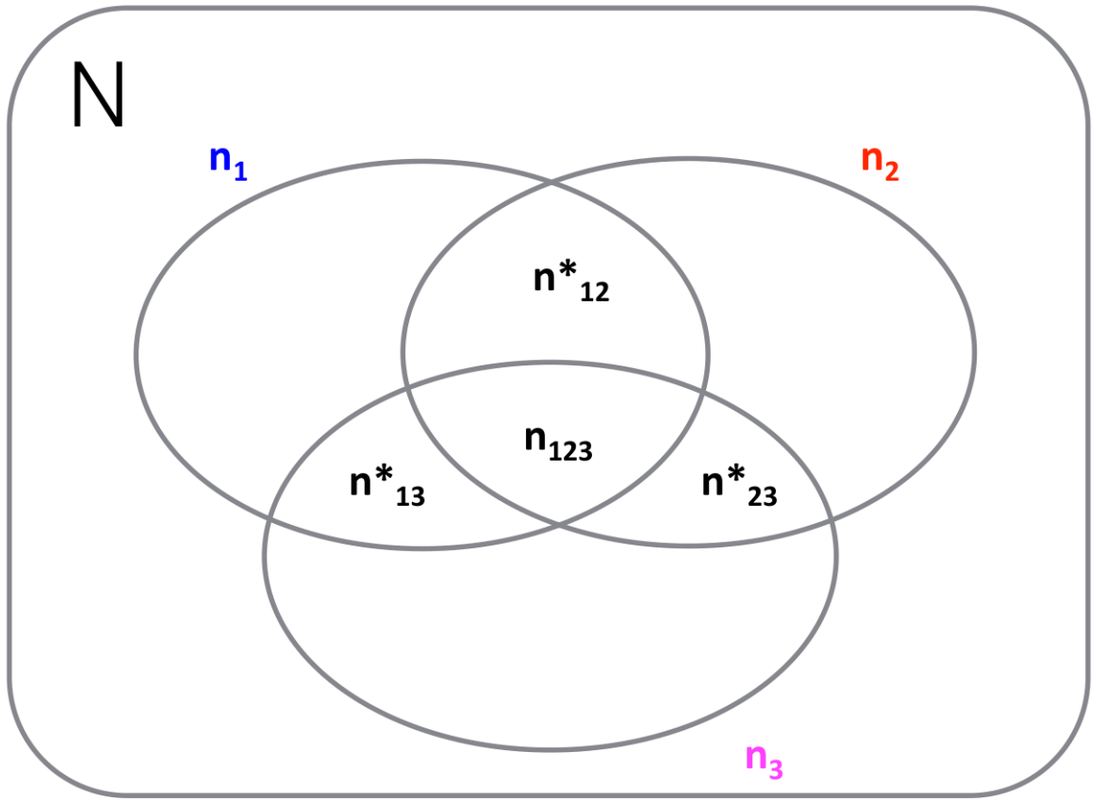


Suspicious



Three events involving three cars

Three-node motifs: statistically validated triangles



Proposition: if random co-occurrence of three subjects, 1,2, and 3, involved in n_1, n_2, and n_3 events, respectively, is assumed in a dataset including N events then

$$p(n^*_{12}, n^*_{13}, n^*_{23} | n_1, n_2, n_3, N) = \frac{\binom{n_1}{n_{12}}\binom{N-n_1}{n_2-n_{12}}\binom{n_{12}}{n_{12}-n^*_{12}}\binom{n_1-n_{12}}{n^*_{13}}\binom{n_2-n_{12}}{n^*_{23}}\binom{N-n_1-n_2+n_{12}}{n_3-n^*_{13}-n^*_{23}-n_{12}+n^*_{12}}}{\binom{N}{n_2}\binom{N}{n_3}}$$

$$\text{p-value} = p\left(n^*_{12} + n^*_{13} + n^*_{23} \geq n^{*,0}_{12} + n^{*,0}_{13} + n^{*,0}_{23}\right)$$

Three-node motifs and antifraud

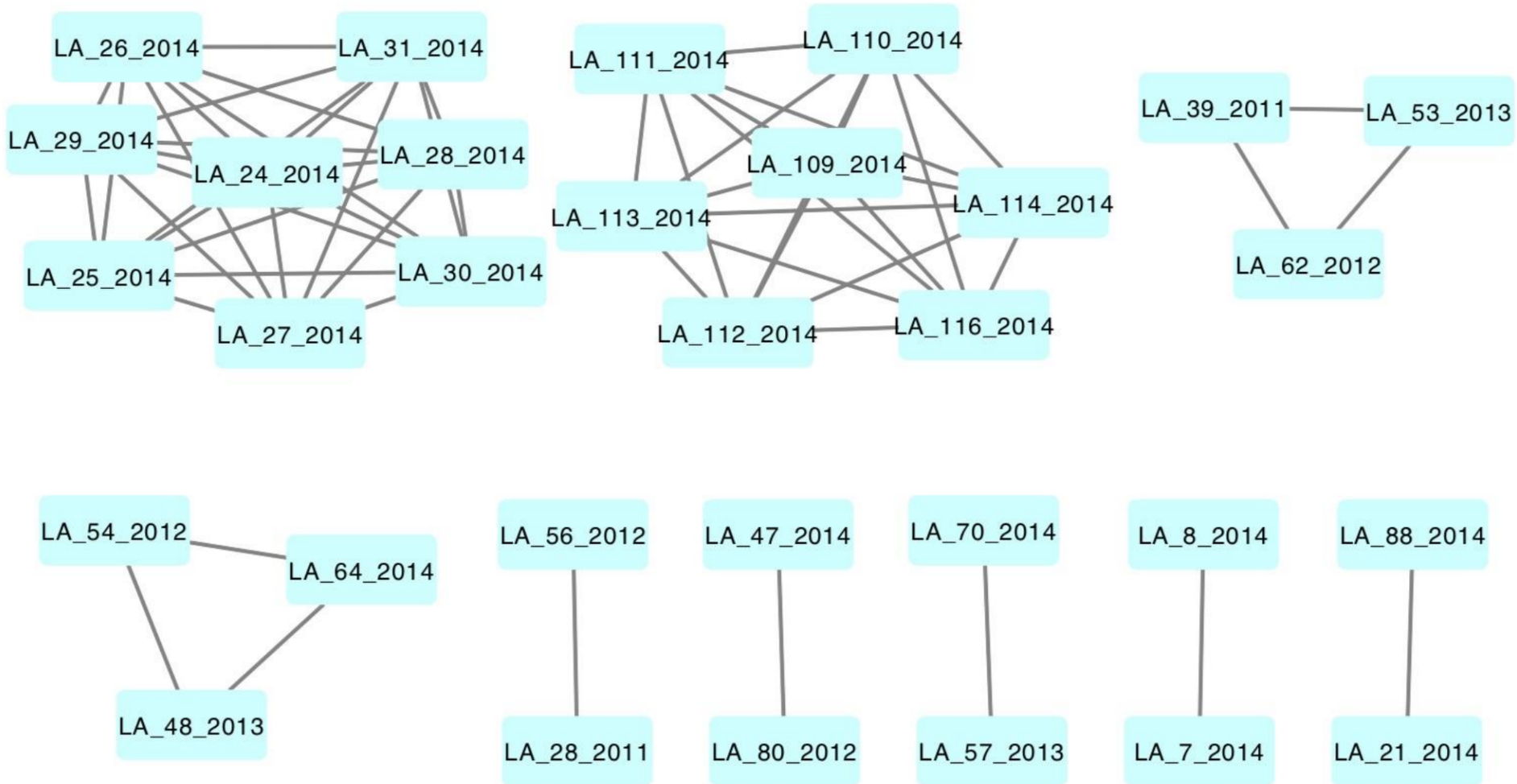
Network of directly involved subjects (no professionals)

- Number of triangles: 162,409
- Number of statistically validated triangles:60,523

Randomly rewired network of directly involved subjects

- Average number of triangles: 18,535
- Average Number of statistically validated triangles: 0.08

Data quality: the statistically validated network of accidents



Final Remarks

1. The **network** of **subjects** and **vehicles** carry different information.

2. Introduced network indicators and IVASS subject indicators carry complementary information, and, therefore, can fruitfully be integrated: the **integrated indicator**.

2. The test on “reported events” and the analysis of several case studies on already identified criminal networks indicate the effectiveness of the overall approach.

2. SVN of accidents for data quality

2. Next step: (a) developing an indicator that involves three-node motifs; (b) evolution of groups of fraudsters