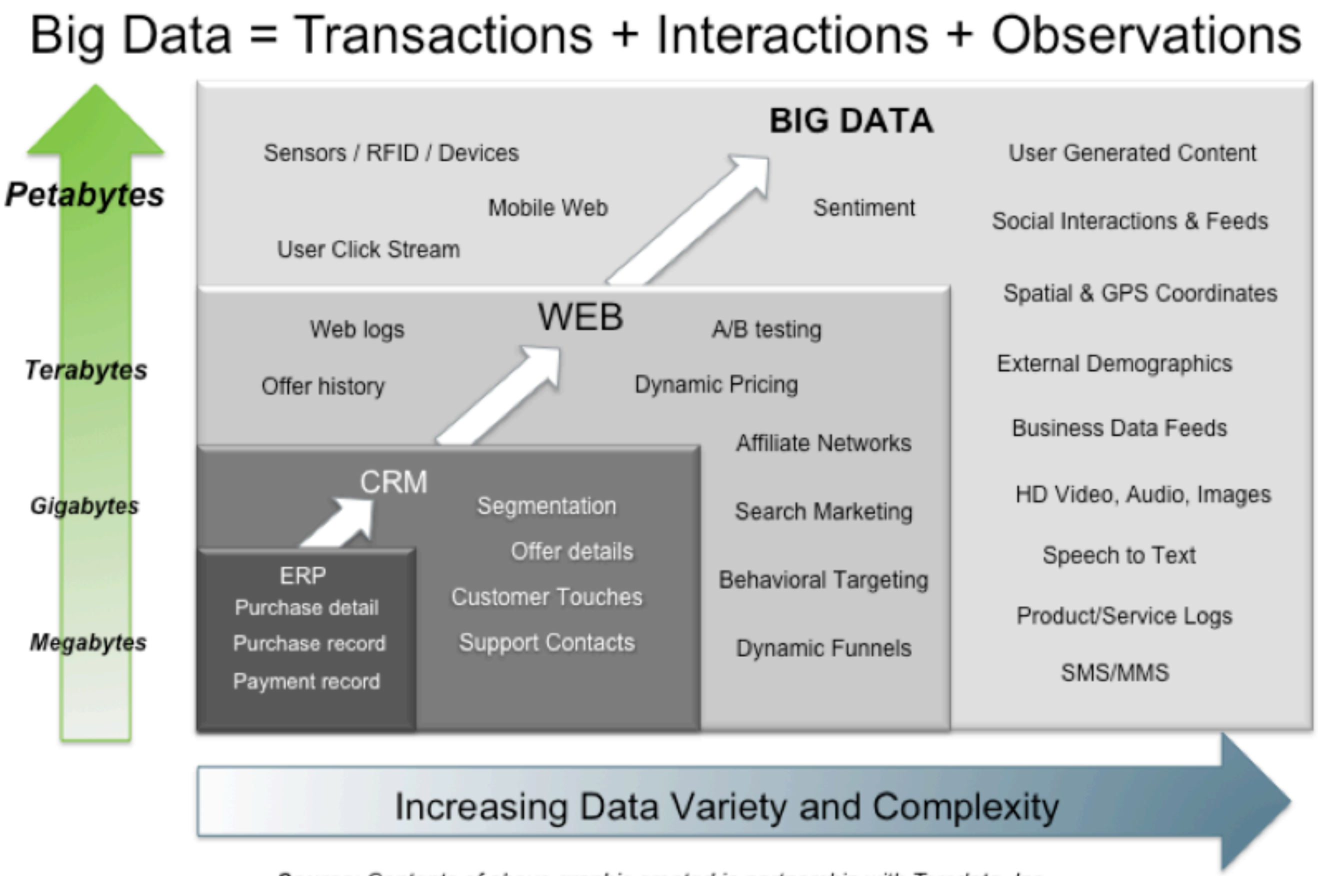
Big Data and Networks for Fraud Detection in the Insurance Sector

Michele Tumminello, Andrea Consiglio
Department of Economics, Management and Statistics
University of Palermo

Summary

- Big Data and the Integrated Antifraud Archive
- Bipartite Networks and statistically validated networks
- Network indicators
- Criminal specialization and network motifs
- Conclusions

Big Data: size does matter



Source: *Contents of above graphic created in partnership with Teradata, Inc.*

Is it just size that matters?

Statistical sample → Inference

Big Data → Filtering

Filtering Big Data of Complex Systems: **issues**

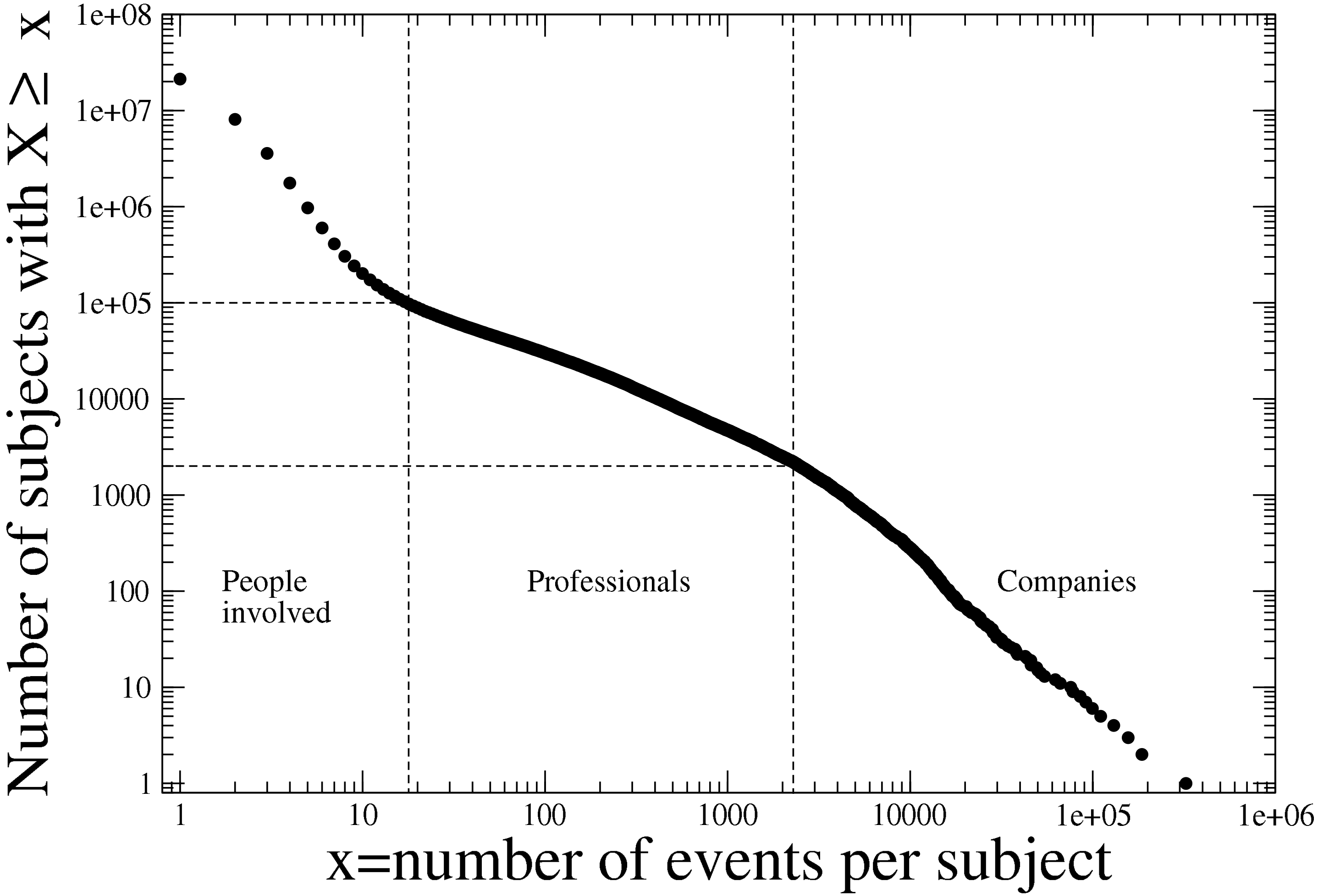
- Heterogeneity
- Integrating and managing information from different sources
- Non-linear interactions and correlations
- Extreme events, Fat tails, Information cascades, Contagion
- Multiple time scales
- Non-stationarity
- Communities & emergent properties
- No controlled experiments
- No reductionism

Big Data: The Integrated Antifraud Archive (AIA)

- Time period: 2011-2016
- About 14 million car accidents
- About 20 million individuals and companies
- About 18 million vehicles

Tumminello M, Consiglio A, **Project**: "*Network analysis and modelling of the integrated anti-fraud database*", funded by the Istituto per la Vigilanza sulle Assicurazioni (**IVASS**), which is the National Agency that supervises the activity of all the insurance companies operating in Italy. Responsible for IVASS: **Farabullini F**

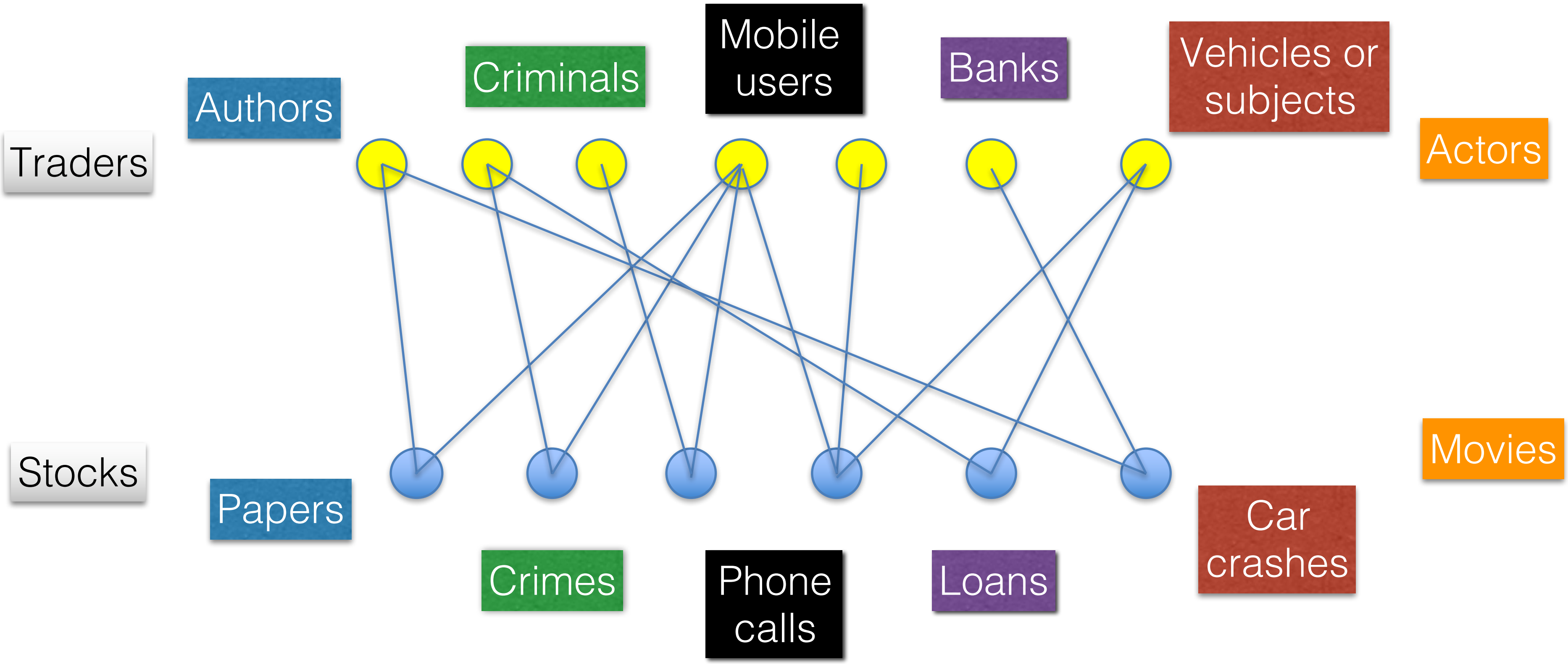
Heterogeneity of subjects



Objectives

- Uncover patterns in the data that suggest fraudulent activity.
- Identify organized groups of perpetrators.

Bipartite networks



NJP
2012
QF
2015

Plos ONE
2013

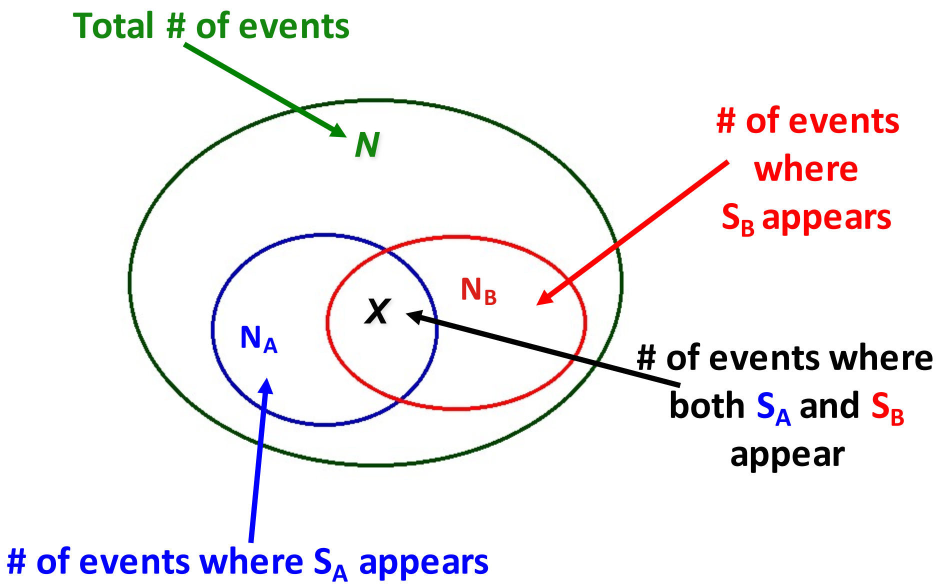
NJP
2015
Sci Rep
2015

QF
2015
JEDC
2015

Plos ONE
2011

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 $\mathbf{S_A}$ and subject $\mathbf{S_B}$ in **X** events against a null hypothesis of random co-occurrence. Suppose that the number of events where $\mathbf{S_A}$ ($\mathbf{S_B}$) appears is $\mathbf{N_A}$ ($\mathbf{N_B}$), whereas the number of events where both $\mathbf{S_A}$ and $\mathbf{S_B}$ appear is ***X***.



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

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***	1,197,055	1,113,389	407.552	**318,876**
Bonferroni network of **vehicles***	209,801	121.253	99,373	**11**

*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)

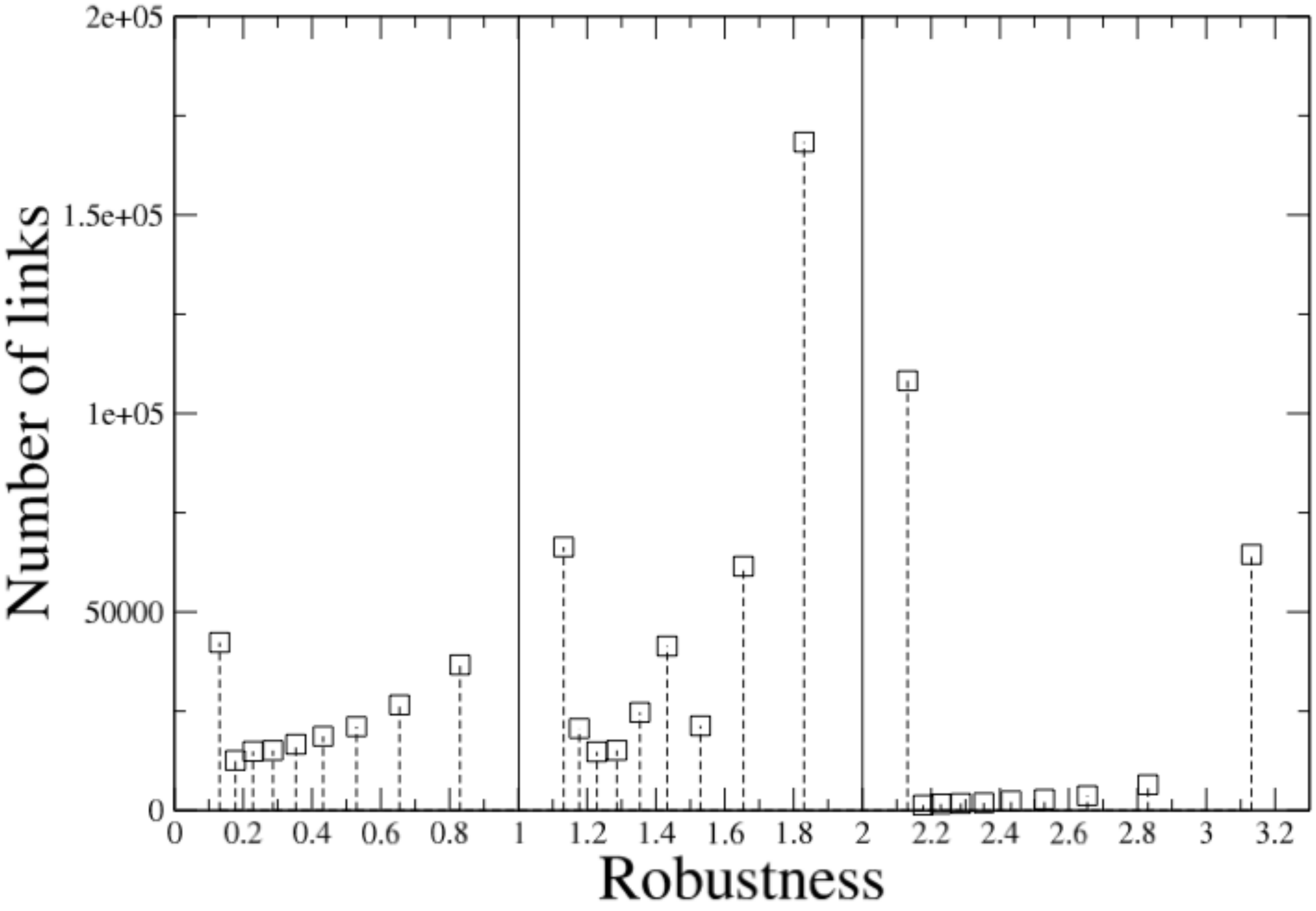
B=bonferroni threshold in the dataset (**B**=1.356e-10 in AIA 10/2016)

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

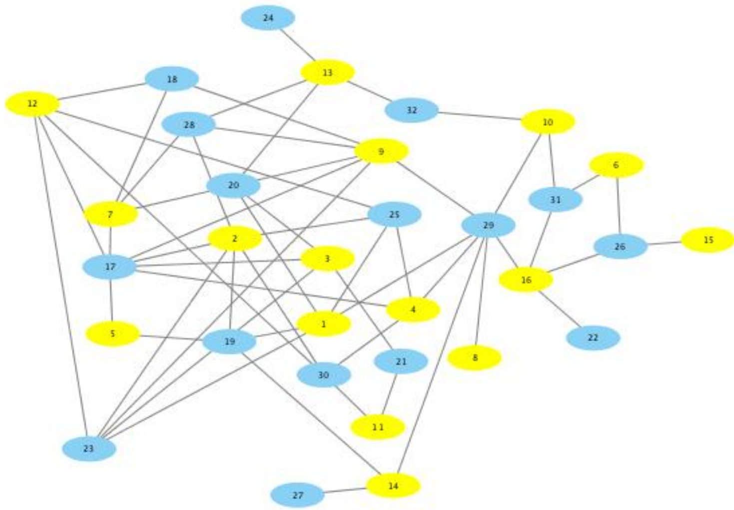
- selecting from the original bipartite network all of the ***event(i)-subject(j)*** pairs such that ***event*(*i*)** contributed to a **link in the SVN between *subject(j)*** and (at least) another subject.

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

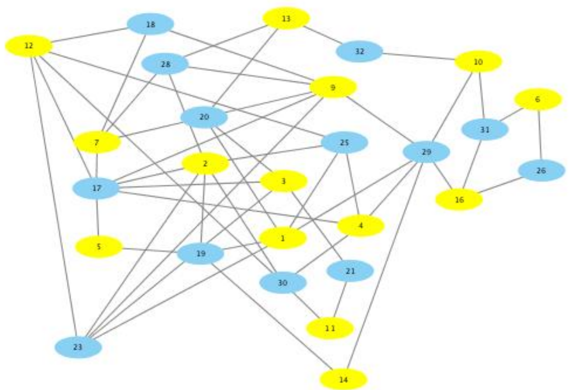
K-H core of a bipartite network

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.

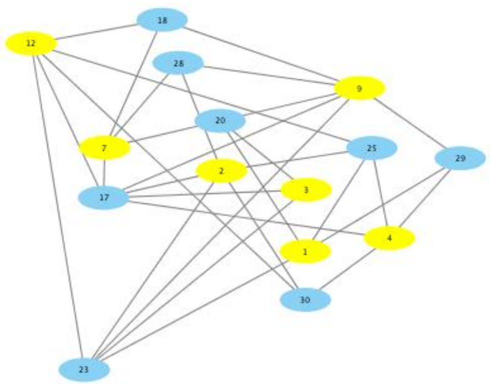
Bipartite network of Kids(blue)-toys(yellow)



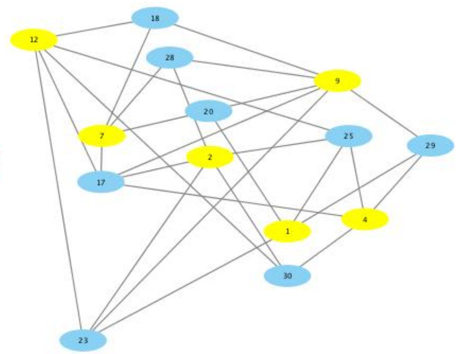
2-2 core



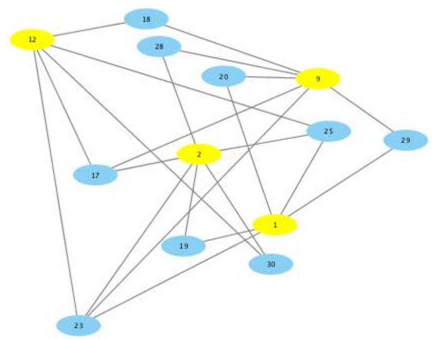
3-3 core



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:

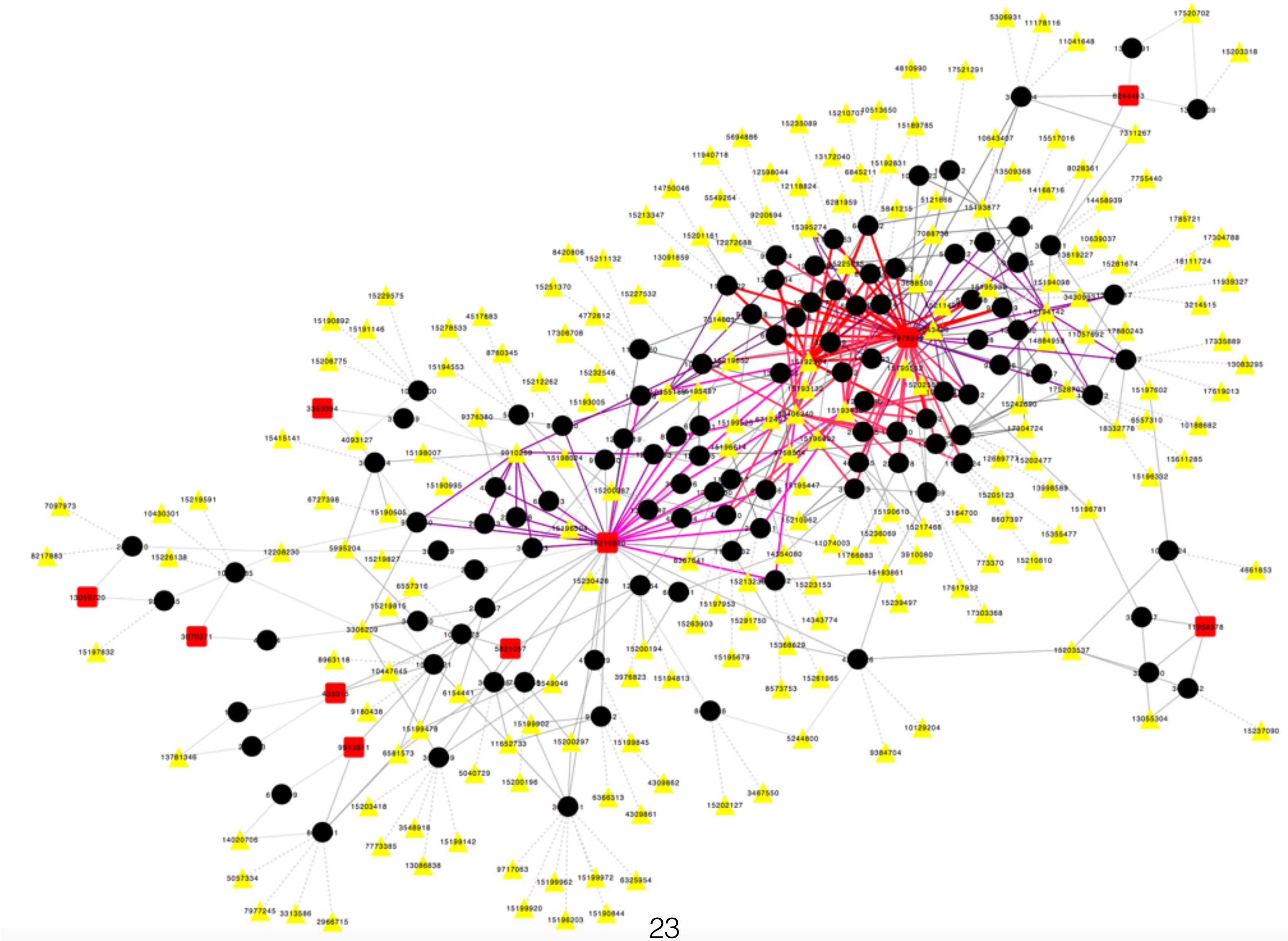
$$KH_s(e,s) = \max(H) \text{ such that } (e,s) \in K - H \text{ core}$$

- **Balanced event-subject indicator**:

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

K-H CORE DECOMPOSITION

of a real statistically validated bipartite subnetwork



Motifs: the heuristics

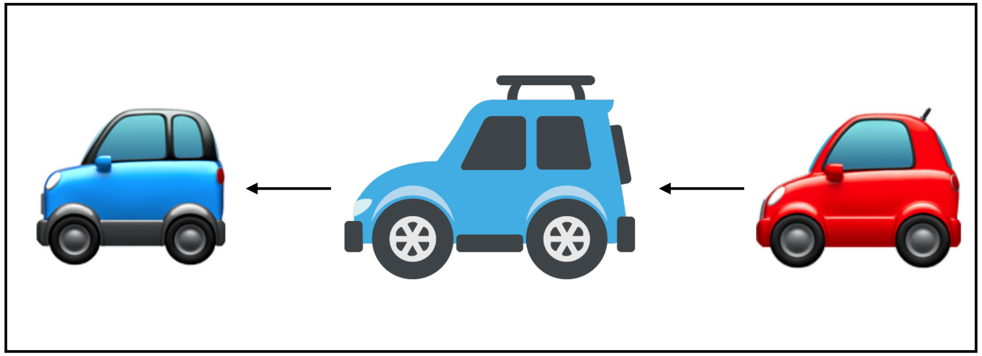
- Criminal specialization
- Some types of crime require cooperation
- Cooperating with a criminal intent requires secrecy and trust

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

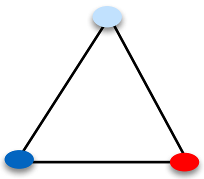
Motifs and anti-fraud

Not suspicious

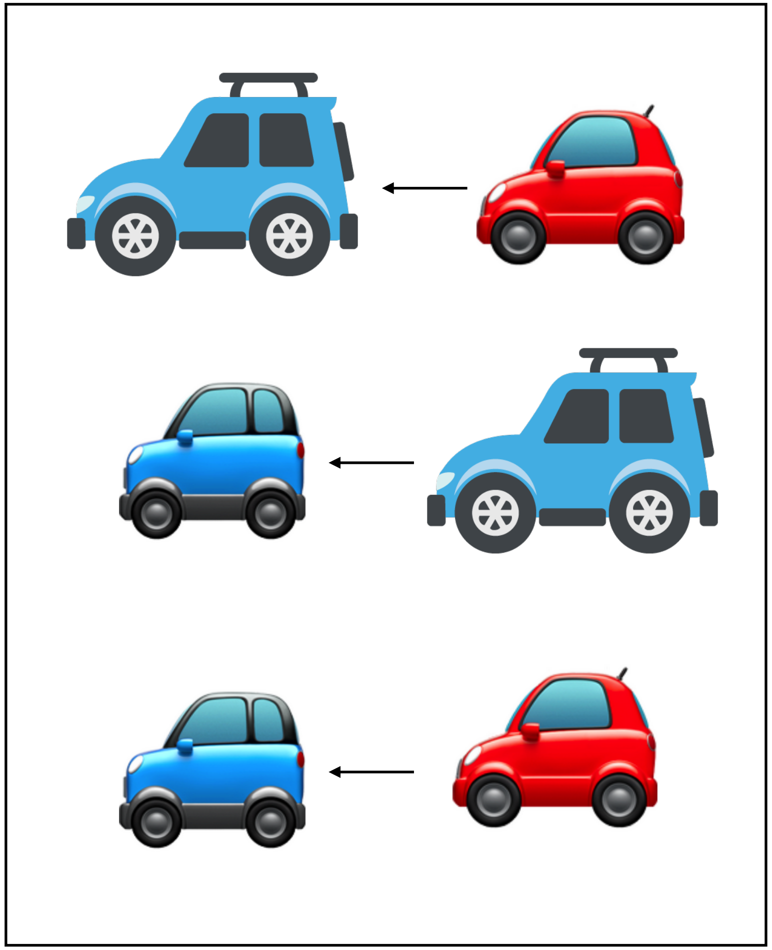


A single event involving three cars

Same projection

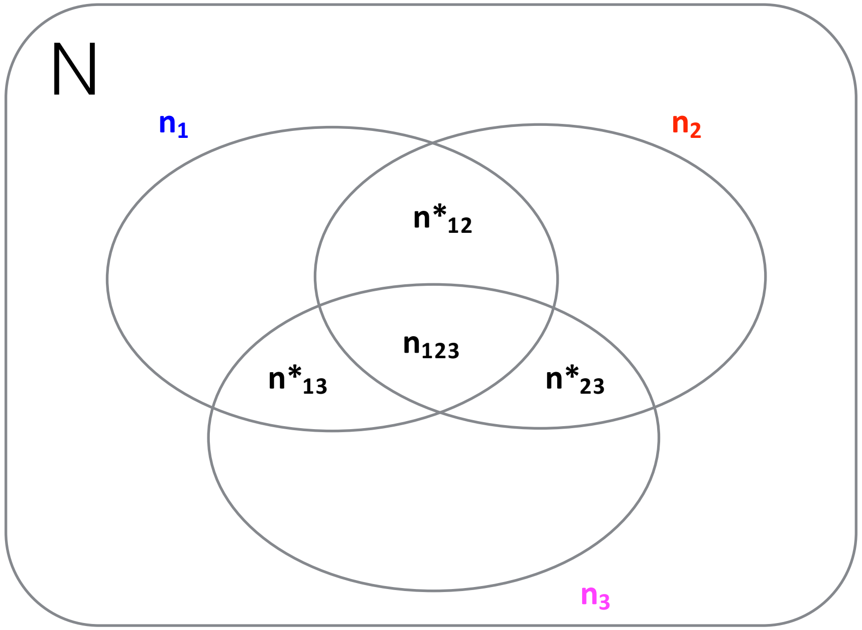


Suspicious



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

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

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

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

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.

3. The test on “claims closed following investigation” and the analysis of a few case studies on already identified criminal networks indicate the effectiveness of the overall approach.

4. Introduced network indicators will be operative by Jan 2018.

5. Next steps: (a) integrating three-node motifs in the SVN (exp. Jun 2017); (b) developing an integrated indicator (exp. end 2018);

Thanks!

Michele Tumminello

Email: michele.tumminello@unipa.it

Alt. Email: michele.tumminello@gmail.com