Network & link analysis

Lovro Šubelj

University of Ljubljana,
Faculty of Computer and Information Science

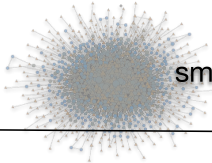
5.2.2014

Network & link analysis

Before 2000

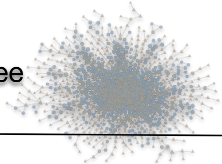
Graph theory

After 1736 — *Present*



small-world networks

scale-free networks

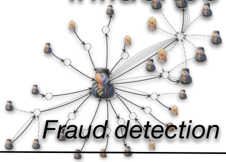


1998 — *1999*

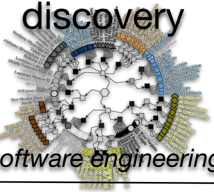
node influence

group discovery

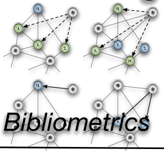
network modeling



Fraud detection



Software engineering

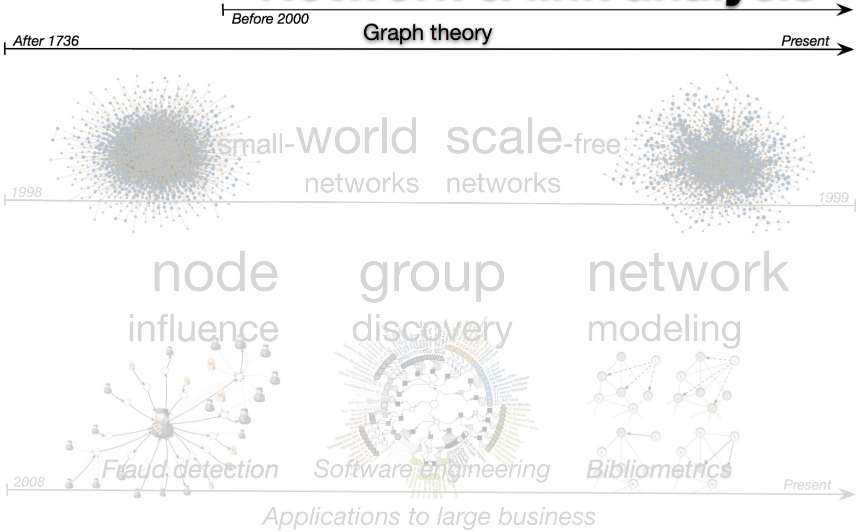


Bibliometrics

2008 — *Present*

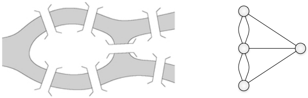
Applications to large business

Network & link analysis



Graph theory

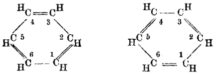
1736 *Königsberg bridge problem* (Euler)



1800s *Travelling salesman problem* (Hamilton)

1845 *Electrical circuit laws* (Kirchhoff)

1857 *Chemical structure* (Kekulé)

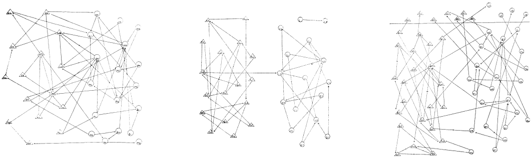


1950s *Operations research* (Dijkstra, Kruskal, Ford)

1959 *Random graphs* (Erdös, Rényi)

Sociometry

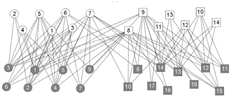
1934 *Sociograms* Moreno (1934)

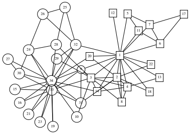


1941 *Southern women* Davis et al. (1941)

1970 *University karate club* Zachary (1977)

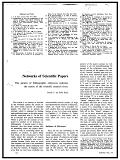
1970s *Social graphs* Granovetter (1973); Freeman (1977, 1979)





Bibliometrics & other

1965 *Scientific papers* Price (1965)

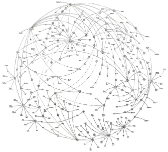

Networks of Scientific Papers

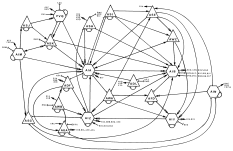
The pattern of bibliographic references indicates the nature of the scientific research front.

Derek J. de Solla Price

SCIENCE CITATION INDEX

1980s *Political scandals* Hobbs and Lombardi (2003)
1986 *Neurology & chemistry* White et al. (1986)
1999 *Transportation* Pelletier (1999)





Network analysis

< 2000 Small graphs (10^2-10^3 nodes)

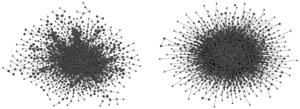
≈ 2000 *Communication networks* (10^5-10^8 nodes)

≈ 2005 *Online social networks* (10^8 nodes)

≈ 2010 *Web graphs* (10^9 nodes)

1998 *Small-world* networks Watts and Strogatz (1998)

1999 *Scale-free* networks Barabási and Albert (1999)



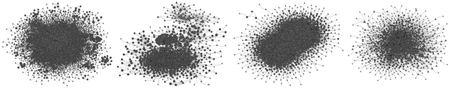
Social, information, biological & technological networks. Newman (2003)

Network & link analysis

network analysis $\rightarrow$ structure & function of networks (*network theory*)

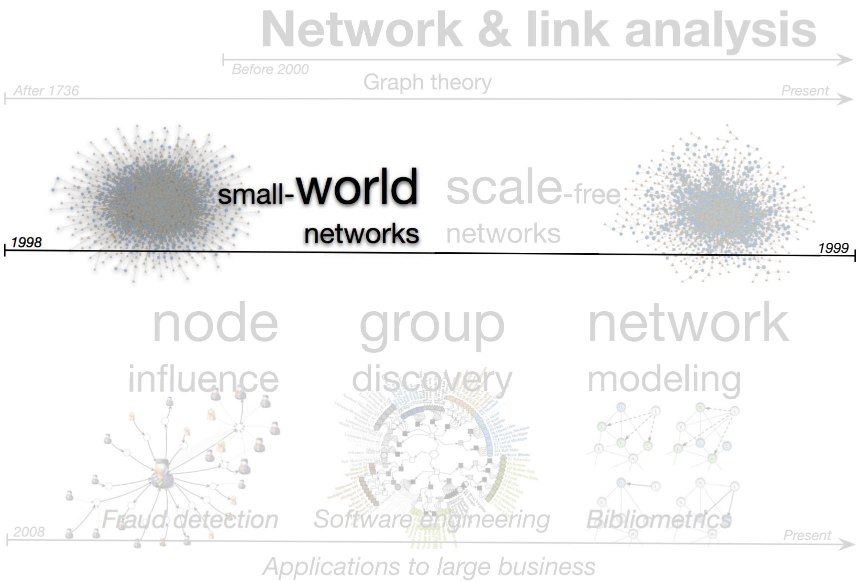
link analysis $\rightarrow$ data mining over nodes using links (*network mining*)

GBDM $\rightarrow$ data mining over graphs (*graph mining*)



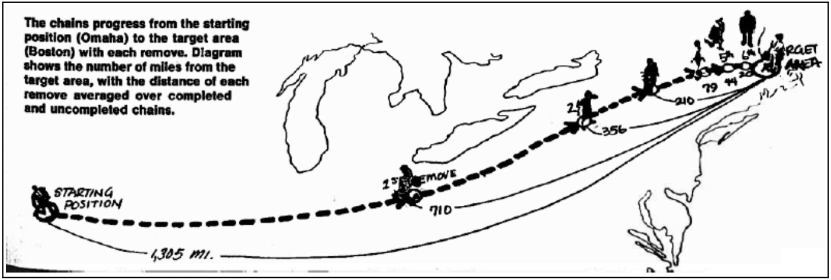
Network analysis & visualization tools:

- *C++* $\rightarrow$ SNAP SNAP (2013)
- *Python* $\rightarrow$ NetworkX Hagberg et al. (2008)
- *Java* $\rightarrow$ JUNG O'Madadhain et al. (2005)
- *Excel* $\rightarrow$ NodeXL Hansen et al. (2010)
- *other* $\rightarrow$ Pajek de Nooy et al. (2005)
  Gephi Bastian et al. (2009)


Network & link analysis
Before 2000
Graph theory
After 1736
Present
small-world
networks
scale-free
networks
1998
1999
node
influence
group
discovery
network
modeling
Fraud detection
Software engineering
Bibliometrics
2008
Present
Applications to large business

Milgram's experiment

Sending a chain letter to a stock-broker in Boston (through friends):
small-world of networks or *6 degrees of separation.* Milgram (1967)



Source: Milgram (1967)

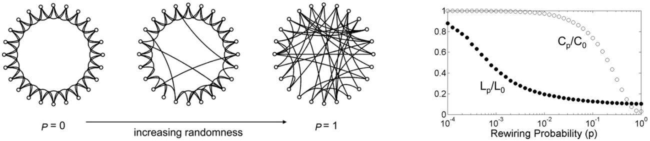
Strength of weak ties & weakness of strong ties. Granovetter (1973)

Small-world graph model

Let k_i be the degree of node i, Δ_i # linked neighbors & d_{ij} the distance between nodes i, j.

$$L = \frac{1}{\binom{n}{2}} \sum_{ij} d_{ij} \qquad C = \frac{1}{n} \sum_{i} \frac{\Delta_i}{\binom{k_i}{2}}$$

Random rewiring of links: Watts and Strogatz (1998)



Source: http://www.frontiersin.org/

Regular graphs have *high L & C*, while random graphs have *low L & C*.

Real-world networks have high C & low L!

Small-world networks

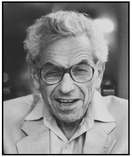
Properties of small-world networks:

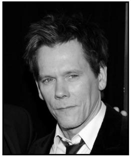
6 degrees of separation Milgram (1967)

7 (4) degrees of separation in e-mail Dodds et al. (2003)

4 degrees of separation on Facebook Backstrom et al. (2012)

Small *Erdös & Bacon numbers* (distances in a collaboration network).





Are small-world networks, e.g., peer-2-peer, also navigable? Kleinberg (2001)

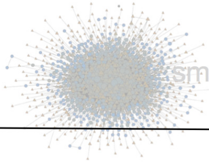
Searchable with a decentralized algorithm in time polynomial in $\mathcal{O}(\log n)$.

Network & link analysis

Before 2000

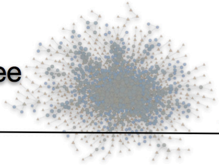
Graph theory

After 1736 — Present



small-world networks

scale-free networks

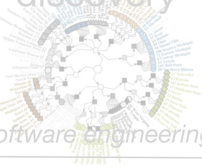


1998 — 1999

node influence

group discovery

network modeling



Fraud detection *Software engineering* *Bibliometrics*

2008 — Present

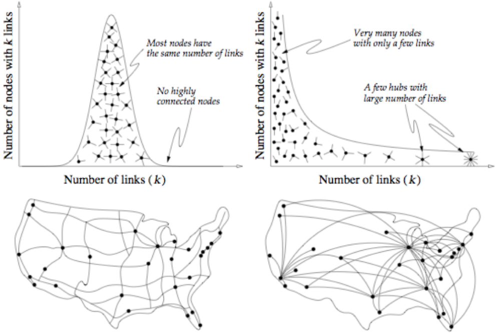
Applications to large business

Normal & power-law distributions

Citations, Internet & the web have *power-law* $P(k)$: Price (1965); Faloutsos et al. (1999)

$$P(k) \sim k^{-\alpha}$$

k_i is the degree of node i & α a power-law exponent, $\alpha > 1$.



Source: Barabási (2002)

Scale-free graph model

Networks with power-law tail of $P(k)$ are called *scale-free*.

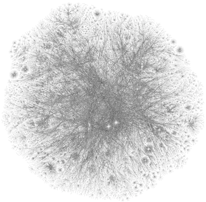
Preferential attachment of nodes: Barabási and Albert (1999)

(1) node i links to a randomly chosen node j with probability p

(2) otherwise, node i links to a node j with probability $\frac{k_j}{\sum k_j}$

$$P(k) = k^{-\alpha} \qquad \alpha = 1 + \frac{1}{1-p}$$

Power-laws arise from *rich get richer phenomena* (cumulative advantage).



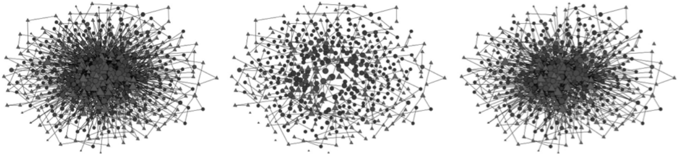
Source: Burch and Cheswick

Scale-free networks

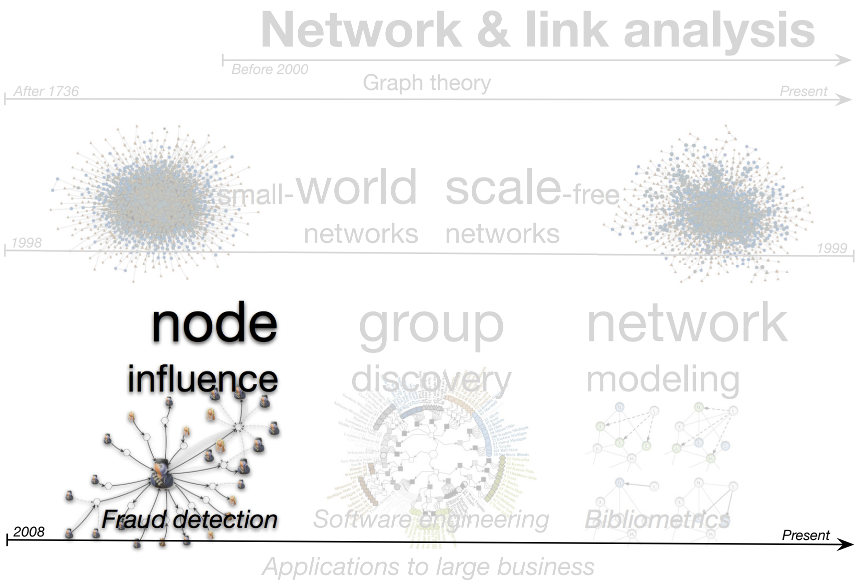
Let n & m be # nodes or links, respectively.

Properties of scale-free networks:

- *sparse* with $m \approx n$ (& not $m \approx n^2$) Del Genio et al. (2011)
- for $\alpha \geq 2$ & $\alpha \geq 3$, the *mean or variance of k_i is infinite* (no scale)
- for $2 < \alpha < 3$, a small infection can *spread in epidemy* Sinha (2011)
- robust against failures & *vulnerable to attacks* Albert et al. (2000)



The Internet with 10% of high-degree or random nodes removed, respectively.


Network & link analysis
Before 2000
Graph theory
After 1736
Present
small-world networks
scale-free networks
1998
1999
node influence
group discovery
network modeling
Fraud detection
Software engineering
Bibliometrics
2008
Present
Applications to large business

Automobile insurance fraud

Staged traffic accidents & fake insurance claims.

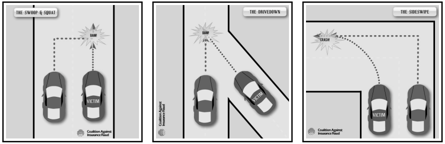
Great risk for other traffic participants (e.g., elders).

$\approx$ 10% outcome for claims only on account of fraud.

$\approx$ 100 million €/year loss for Slovenia (population 2 million).

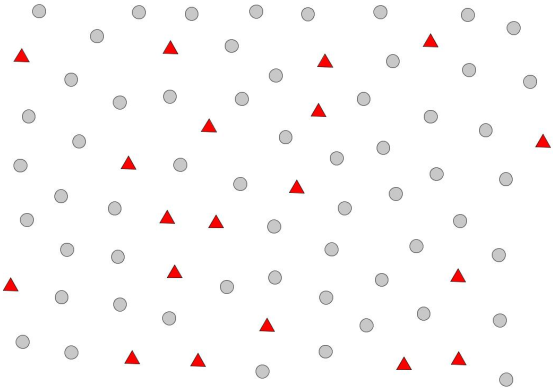
Particularly interesting are groups of collaborating fraudsters.

Design of expert system applicable in practice (e.g., reports).



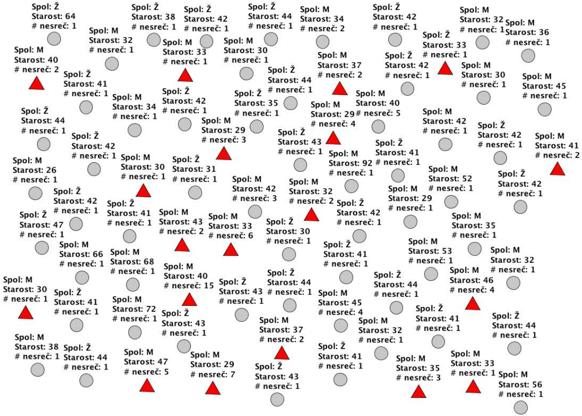
Source: http://www.insurancefraud.org/

State-of-the-art in fraud detection



State-of-the-art in fraud detection

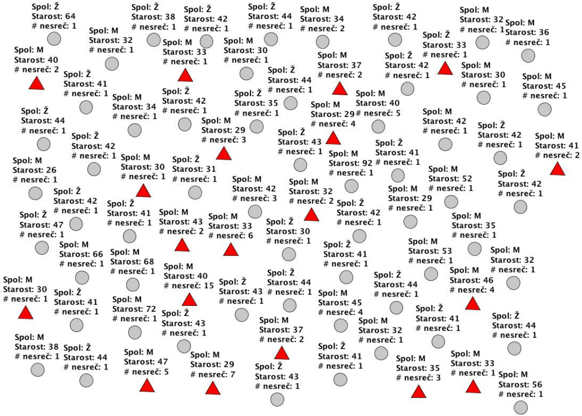
Statistics, machine learning, data mining (labeled data) or experts.



State-of-the-art in fraud detection

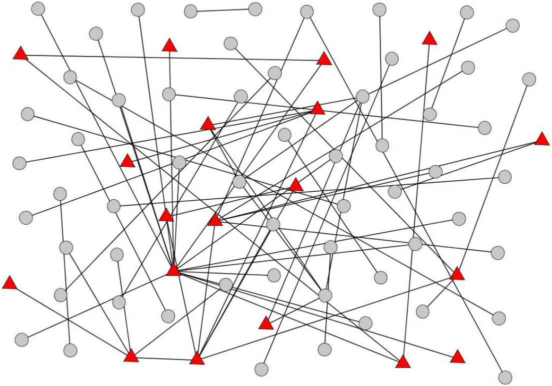
Statistics, machine learning, data mining (labeled data) or experts.

No differences in practice & much fraud is undetected.



Social network analysis

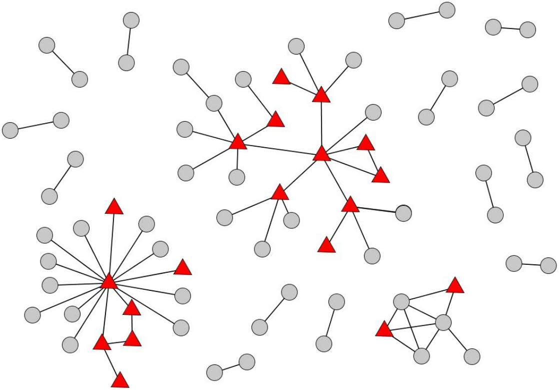
Traffic participants are linked to form a social network.



Social network analysis

Traffic participants are linked to form a social network.

Fraudsters can be detected already with a naked eye.



Expert system for fraud detection

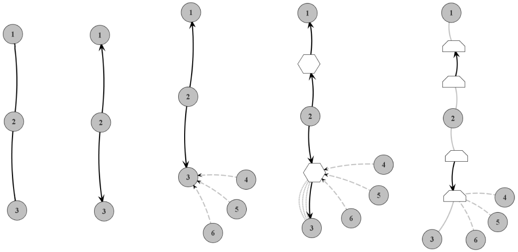
Four-phase fraud detection system: Šubelj et al. (2011, 2009)

(1) *Projection* to a social network

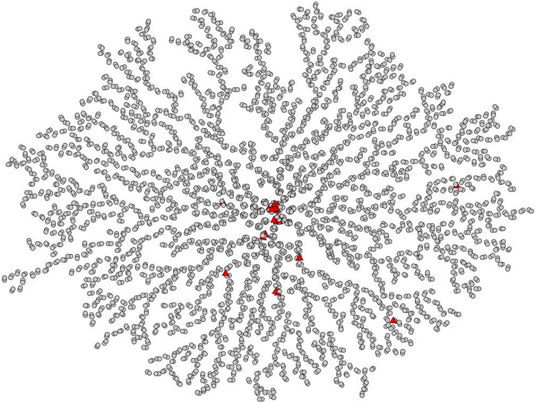
(2) Detection of *suspicious groups*

(3) Detection of *suspicious participants*

(4) *Representation* of the results

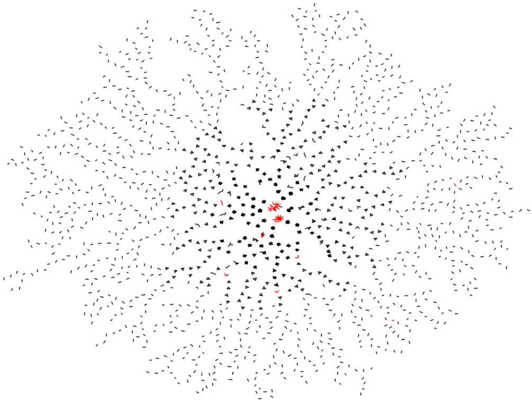


Suspicious groups detection



Suspicious groups detection

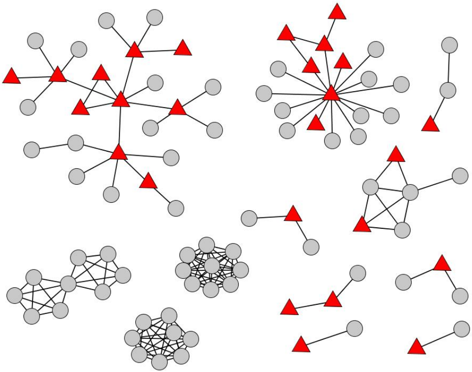
Network decomposes into several connected components.



Suspicious groups detection

Network decomposes into several connected components.

Indicators of common features of fraudulent components.

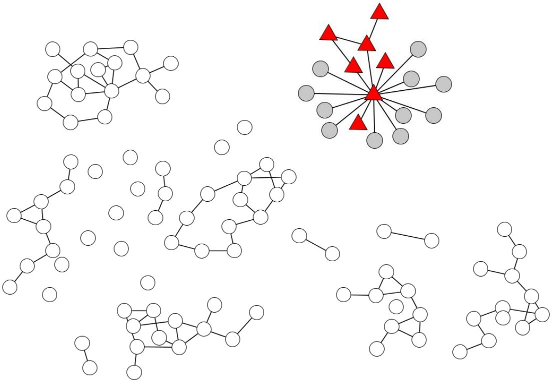


Suspicious groups detection

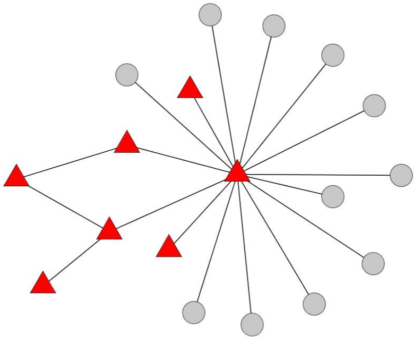
Network decomposes into several connected components.

Indicators of common features of fraudulent components.

Suspicious groups (i.e., components) are detected by simulation.

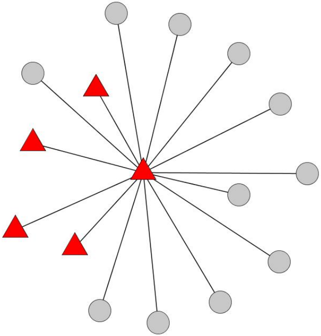


Suspicious participants detection



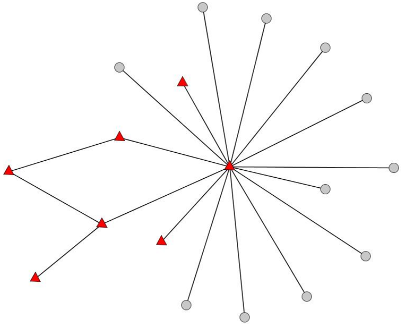
Suspicious participants detection

Birds of a feather flock together in two-mode networks.



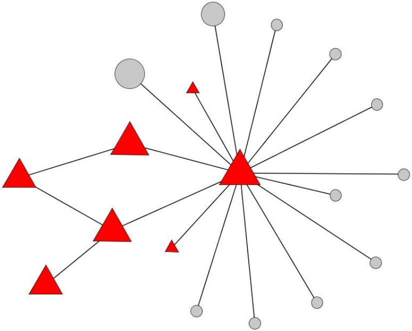
Suspicious participants detection

Birds of a feather flock together in two-mode networks.



Suspicious participants detection

Birds of a feather flock together in two-mode networks.
Propagation of suspicion over the network overcomes locality.
Suspicious participants are detected by a link analysis algorithm...



Link analysis algorithm

Iterative assessment algorithm (with Laplace smoothing): Šubelj et al. (2011)

$$s_i = \frac{1 + k/k_i}{2} \left(f_i \sum_{j \in \Gamma_i} f_{ij} \cdot s_j \right)$$

s_i is score of node i, f_i its factor, k_i its degree & k the mean degree, $f : i \to [0, \infty)$.

$$f_i = \prod_k f_i^k \qquad f_i^k = \begin{cases} 1/(1 - F_i^k) & F_i^k \geq 0 \\ 1 + F_i^k & F_i^k < 0 \end{cases}$$

F are suspicion factors set by an expert, $F : i \to (-1, 1)$.

HITS Kleinberg (1999) & PageRank Page (2001) are *not directly applicable* here.

Social network centrality

Degree centrality: see Scott (2000)

$$c_i = \frac{k_i}{n-1}$$

Closeness centrality: Freeman (1979)

$$c_i = \frac{1}{n-1}\sum_j d_{ij}$$

Betweenness centrality: Freeman (1977)

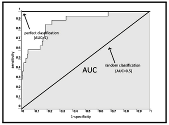
$$c_i = \frac{1}{\binom{n}{2}}\sum_{j,k} \sigma_{jk}(i)/\sigma_{jk}$$

Eigenvector centrality: Bonacich (1987)

$$c_i = \frac{1}{\kappa}\sum_{j \in \Gamma_i} c_j$$

d_{ij} is the distance & σ_{ij} the # geodesics between nodes i, j. κ is the leading eigenvector.

Traffic accidents in Slovenia 1999–2008



Area under the ROC for *groups & participants ranking*: Šubelj et al. (2011)

	Cover	L_1	BC	**MAJOR**	RIDIT	**PRIDIT**
All	0.6019	0.6386	0.6774	**0.7946**	0.6843	0.7114
Suspicious	0.6119	0.8494	0.8549	0.8507	0.9221	**0.9228**

						IAA algorithm		
	ML/DM	DC	CC	BC	EC	**No F**	Raw *F*	**Expert F**
All	/	0.7428	0.8138	0.6401	0.7300	**0.8188**	0.8435	**0.8787**
Suspicious	≈ 0.86	0.8597	0.8158	0.6541	0.8581	**0.8942**	0.9086	**0.9228**

Awards & other

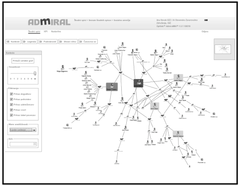
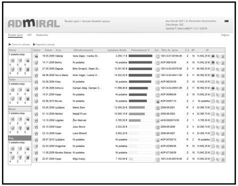
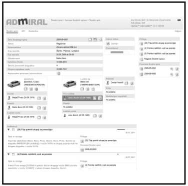
Publication awards:

journal exceptional work by Slovenian Research Agency! Šubelj et al. (2011)

thesis Prešeren award by Faculty of Computer Science Šubelj (2008)

conference best student paper award at DSI '09 Šubelj et al. (2009)

Optilab offers tool *Admiral* adopted by Slovenian Insurance Association.

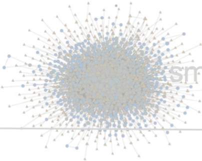
Network & link analysis

Before 2000

Graph theory

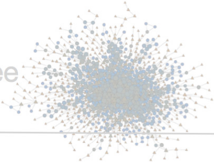
After 1736

Present



small-world networks

scale-free networks



1998

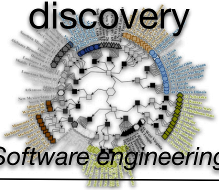
1999

node influence

group discovery

network modeling

Fraud detection



Software engineering

Bibliometrics

2008

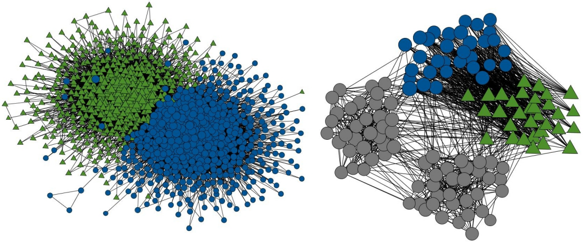
Present

Applications to large business

Groups in real-world networks

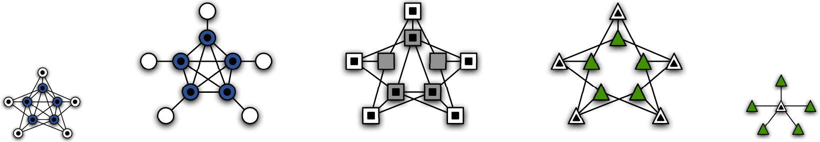
community densely linked nodes that are sparsely linked between
(or dense groups of sparse graphs) Girvan and Newman (2002)

module nodes linked to similar other nodes Newman and Leicht (2007)
(or groups with similar linking pattern) Šubelj and Bajec (2012b)



Group type formalism

Let S be a group (filled) & T its linking pattern (marked). Šubelj et al. (2013a)



Community ($S = T$) *Mixture* ($S \approx T$) *Module* ($S \neq T$)

Let $\tau_{S,T}$ be a parameter of group S & its pattern T.

$$\tau_{S,T} = \frac{|S \cap T|}{|S \cup T|}$$

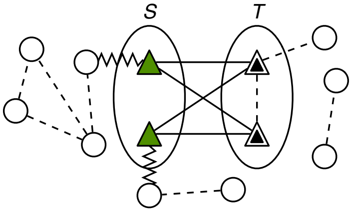
$\tau = 1$ for communities, $\tau \approx \frac{1}{2}$ for mixtures & $\tau = 0$ for modules...

Group quality criterion

Let $L_{S,T}$ be a number of links between S & T.

$$W_{S,T} = \dots \left(\frac{L_{S,T}}{|S||T|} - \frac{L_{S,T^C}}{|S||T^C|} \right)$$ Šubelj et al. (2013a)

A local asymmetric criterion that *favors links* in (S, T) & *penalizes for links* in (S, T^C). Consistent with wide class of models for $S = T$. Zhao et al. (2011)



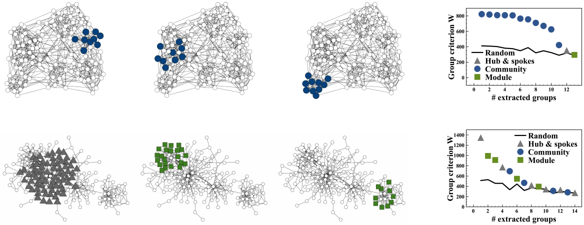
Group discovery by extraction

Sequential group extraction: Šubelj et al. (2013a) & Zhao et al. (2011)

(1) Find S & T that optimize W (tabu search)

(2) Extract *only links between* S & T (& isolated nodes)

(–) Repeat until W larger than at random (by simulation)

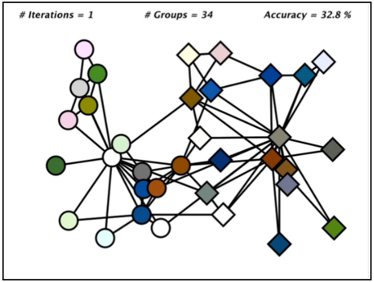


Group detection by propagation (intermezzo)

Propagation group detection: Raghavan et al. (2007)

$$s_i = \operatorname*{argmax}_s \sum_{j \in \Gamma_i} \delta(s_j, s)$$

s_i is (group) label of node i & Γ_i are its neighbors.



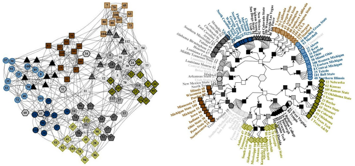
Group detection by propagation (II)

performance	Diffusion propagation Šubelj and Bajec (2011b)
robustness	Balanced propagation Šubelj and Bajec (2011a)
generality	General propagation Šubelj and Bajec (2012b)

Hierarchical group detection: Šubelj and Bajec (2014)

$$s_i = \operatorname*{argmax}_s \left(\overbrace{\tau_s \cdot \sum_{j \in \Gamma_i} \ldots \delta(s_j, s)}^{\text{Community detection}} + \overbrace{(1 - \tau_s) \cdot \sum_{\substack{j \in \Gamma_i \\ k \in \Gamma_j \setminus \Gamma_i}} \ldots \delta(s_k, s)}^{\text{Module detection}} \right)$$

The algorithm is *at least comparable* to the state-of-the-art! Šubelj (2013)



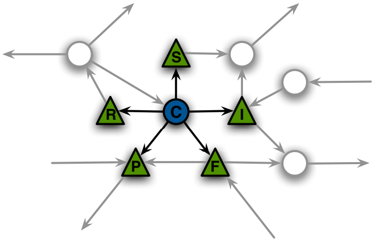
Software networks

Class dependency software networks: Šubelj and Bajec (2011c)

nodes → *classes* of an object-oriented software project

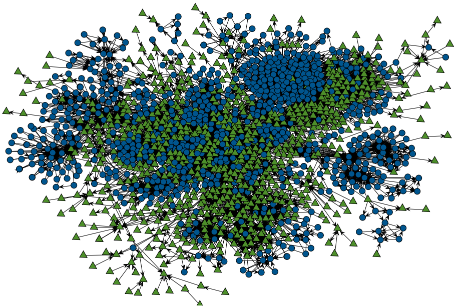
links → *dependencies* between classes (e.g., inheritance)

```
class C extends S implements I {

      F field;

      public C() { ... }

      void foo(P parameter) { ... }

      private R bar() { ... }

}
```



Structure of software networks

Software networks are similar to other real-world networks. Valverde et al. (2002)

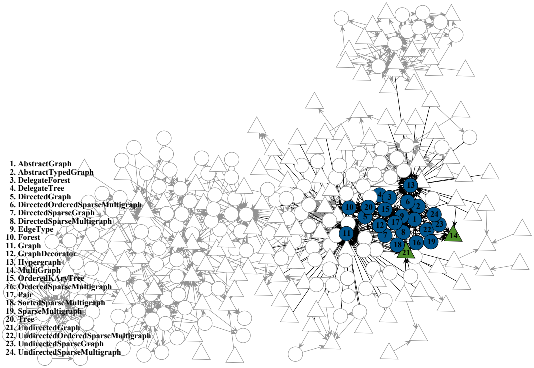


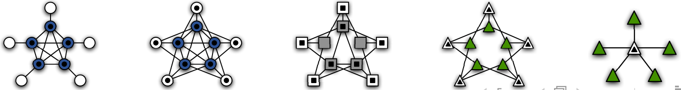
software networks = Šubelj et al. (2013b)

= **dense social network structure** + **sparse Internet topology**

Communities in software networks

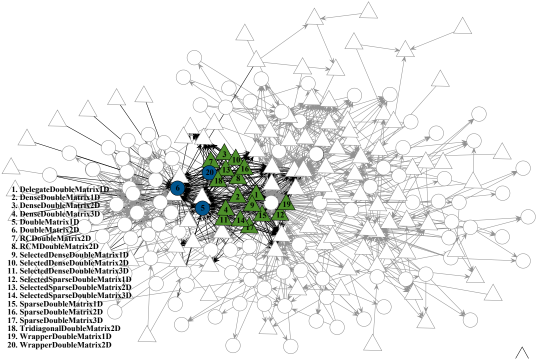
Communities are *core classes* of the software project. Šubelj and Bajec (2011c)

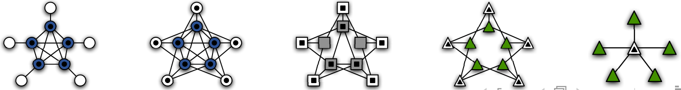




Modules in software networks

Modules are classes with the *same functionality*. Šubelj and Bajec (2012b)





Software engineering

Accuracy of *class package* prediction: Šubelj et al. (2013b)

Software	# Classes	# Categories	Neighbors Γ	**Groups** *S*	Network *N*	*Baseline*	*Random*
JBullet	107	11	72.0%	**75.7%**	64.5%	*28.0%*	*8.6%*
colt	154	16	58.4%	**73.4%**	55.2%	*22.7%*	*5.9%*
JUNG	237	31	72.2%	**74.2%**	65.0%	*11.4%*	*3.3%*
Lucene	1335	178	47.1%	**49.2%**	43.7%	*6.4%*	*0.6%*

Accuracy of *high-level class package* prediction:

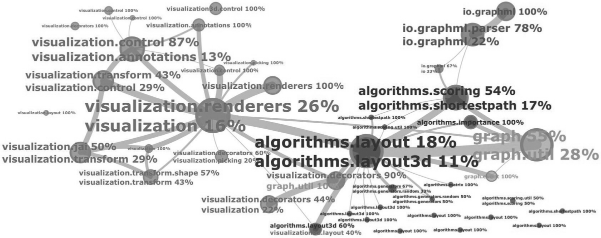
Software	# Classes	# Categories	Neighbors Γ	**Groups** *S*	Network *N*	*Baseline*	*Random*
JBullet	107	5	**84.6%**	**85.0%**	78.5%	*64.5%*	*20.4%*
colt	154	10	**86.4%**	83.8%	69.5%	*39.0%*	*9.7%*
JUNG	237	5	89.1%	**90.5%**	**91.1%**	*44.3%*	*20.3%*
Lucene	1335	15	85.5%	**90.8%**	85.0%	*28.2%*	*6.6%*

Accuracy of *class type, version, author* prediction:

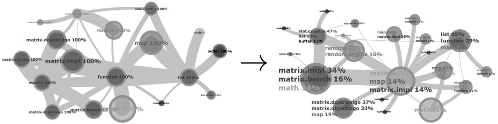
Setting	# Categories	Neighbors Γ	**Groups** *S*	Network *N*	*Baseline*	*Random*
Class type	2	65.0%	**85.2%**	**84.8%**	**84.4%**	*49.9%*
Class version	9	67.7%	**72.8%**	66.2%	*44.3%*	*11.2%*
Class author	11	**71.6%**	**71.0%**	**70.9%**	*44.3%*	*9.2%*

Software engineering (II)

High-level abstraction of a software system: Šubelj and Bajec (2012a)



Reorganization of software packages (modular or functional):



$Q = -0.0262$ $\qquad$ $Q = 0.5020$

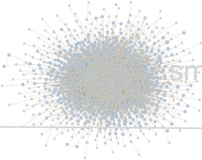
Network & link analysis

Before 2000

Graph theory

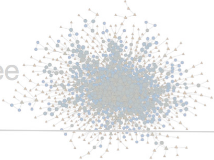
After 1736

Present



small-world networks

scale-free networks



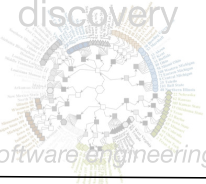
1998

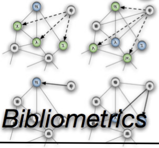
1999

node influence

group discovery

network modeling





Fraud detection

Software engineering

Bibliometrics

2008

Present

Applications to large business

Social & non-social networks

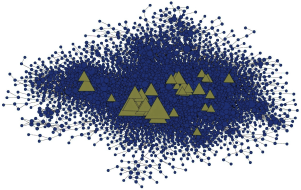
Real-world networks are small-world, scale-free, shrink & densify.

Degree mixing (correlations of degrees at links' ends): Newman (2002)

Social networks $\rightarrow$ *assortative* Newman and Park (2003)

Non-social networks $\rightarrow$ *disassortative* Šubelj and Bajec (2012b)

Citation networks $\rightarrow$ *neither*



Forest fire graph model

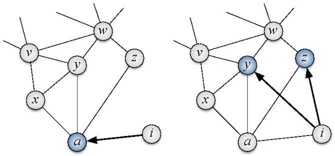
Sequential node inclusion: Leskovec et al. (2007)

(1) node i selects ambassador a & links to a

(2) node i selects neighbors $a_1, \ldots, a_{x_p}$ & links to a_i

(3) $a_1, \ldots, a_{x_p}$ are taken as ambassadors

p is burning probability & $x_p \sim \mathrm{G}(\frac{p}{1-p})$.

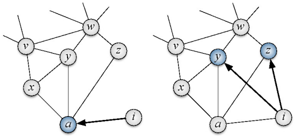


Generated graphs are small-world, scale-free & *degree assortative*.

Model of citation networks

Author citation dynamics:

(1) author selects a paper & cites it
(2) author selects its references & cites them
(3) references are taken for consideration



Authors should read all papers they cite (& vice-versa).

Only $\approx 20\%$ cited papers are actually read. Simkin and Roychowdhury (2003)

Authors read & cite papers due to independent processes.

Citation graph model

Sequential node inclusion: Šubelj and Bajec (2013)

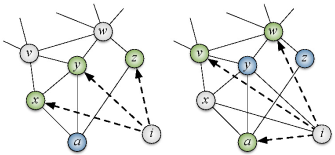
(1) node i selects ambassador a

(2) node i selects neighbors $a_1, \dots, a_{x_p}$

node i selects neighbors $l_1, \dots, l_{x_q}$ & links to l_i

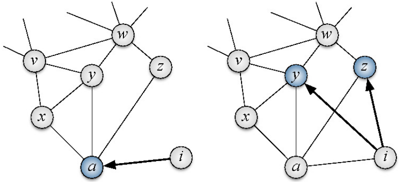
(3) $a_1, \dots, a_{x_p}$ are taken as ambassadors

q is linking probability & $x_q \sim \mathrm{G}(\frac{q}{1-q})$.

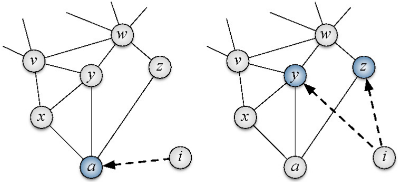


Generated graphs are small-world, scale-free & *degree (dis)assortative*.

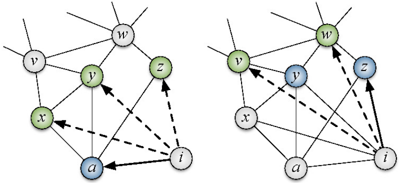
Alternative graph models



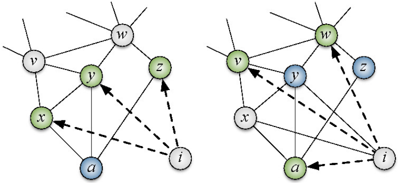
Forest fire model Leskovec et al. (2007)



Butterfly model McGlohon et al. (2008)



Copying model Krapivsky and Redner (2005)

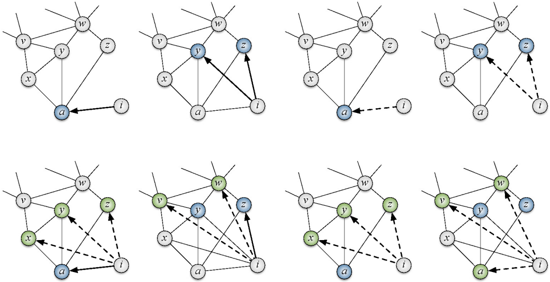


Citation model Šubelj and Bajec (2013)

Analysis of graph models

Let A be the set of ambassadors & L the set of linked nodes.

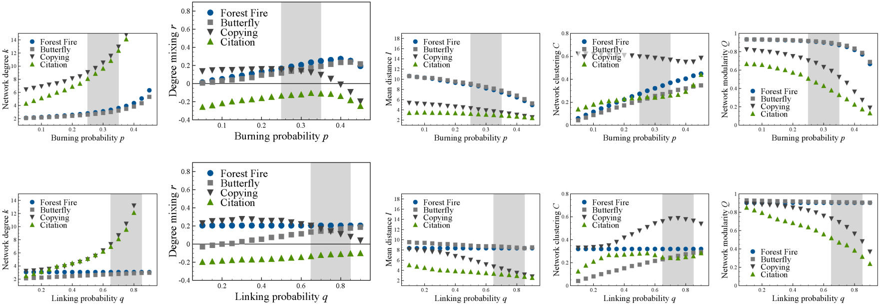
Forest fire model $\rightarrow A = L$
Butterfly model $\rightarrow A \supseteq L$
Copying model $\rightarrow A \subseteq L$
Citation model $\rightarrow A$ & L *arbitrary*



Comparison of graph models

All models generate *small-world & scale-free* graphs with high *modularity*.

Shaded regions show likely parameter values. Laurienti et al. (2011)



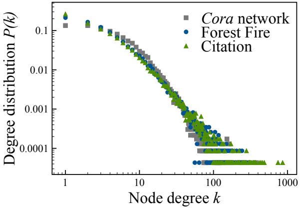
Only citation model realizes *degree disassortative* graphs! Šubelj and Bajec (2013)

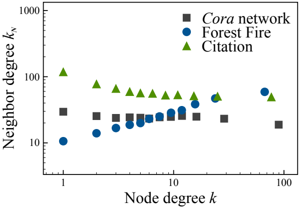
Cora citation network

Cora dataset of computer science papers from the web. McCallum et al. (2000)

	p	q	# Nodes	# Links	Degree k	**Mixing r**
Forest fire model	0.46	/	23166	88828	7.669	**0.211**
Citation model	0.37	0.59		89888	7.760	−**0.047**
Cora network			23166	89157	7.697	−**0.055**

Citation model reproduces *degree disassortativity* of *Cora* network.



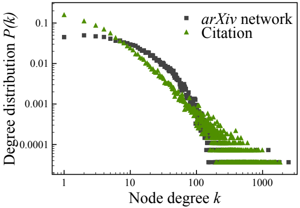


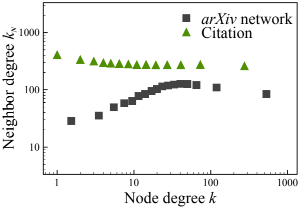
arXiv citation network

High energy particle physics preprints from *arXiv* server. KDD (2003)

	p	q	# Nodes	# Links	Degree k	**Mixing r**
Citation model	0.46	0.67	27400	350699	25.598	$-$**0.068**
arXiv network			27400	352021	25.695	$-$**0.030**

Directed *arXiv* network is modeled with an *undirected graph*...





Biblio- & scientometrics

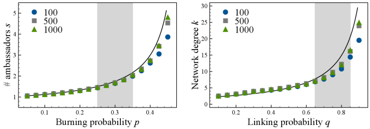
% *Papers read* relative to # *papers cited* is

$$\approx 2s/k$$

$$s \leq \frac{1-p}{1-2p} \qquad k \leq \frac{2ps}{1-q-(1-q)^{s+1}}$$

s is # ambassadors selected by a node.

	p	q	# Cited	# Read	**% Read**
Cora network	0.37	0.59	3.85	2.54	**66%**
arXiv network	0.46	0.67	12.85	6.30	**49%**



Biblio- & scientometrics (II)

$\#$ *Papers cited or citing* is $\approx k/2$

% *Papers read* relative to $\#$ *papers cited* is $\approx 2(1-p)/(k-2kp)$

% *Papers cited* relative to $\#$ *papers read* is $\approx q/(1-q)$

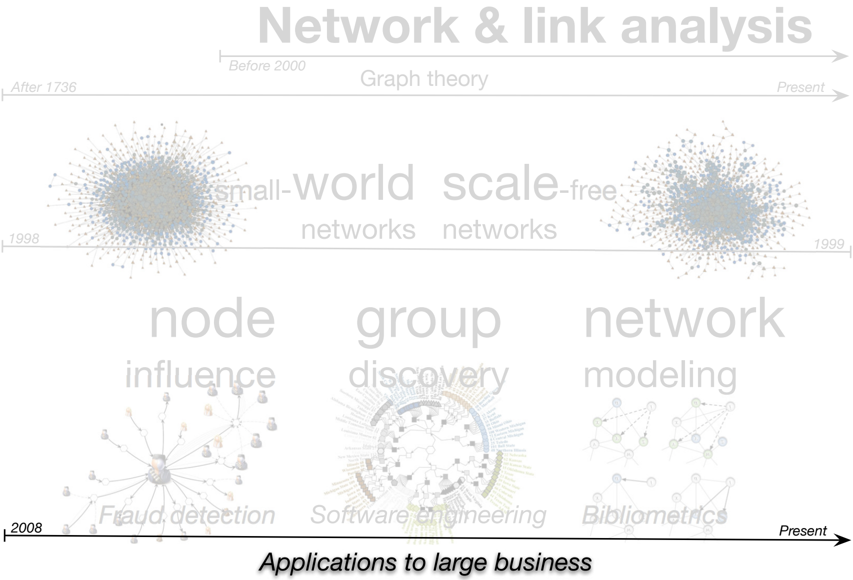
Directed citation model (ongoing work):

burning process $\rightarrow$ probabilities p_{fwd} & p_{bck}

linking process $\rightarrow$ probabilities q & q_{amb}

Citation dynamics of scientific fields from WoS (future work).

THOMSON REUTERS

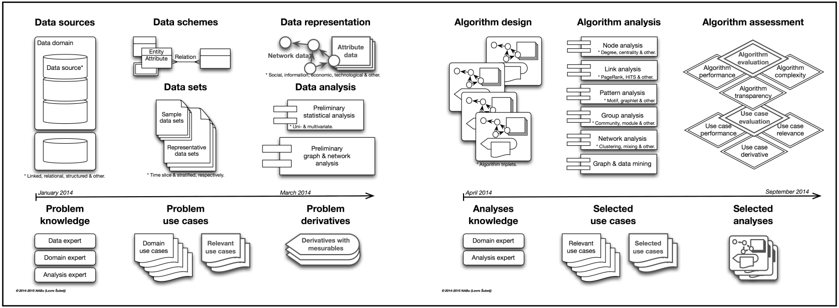

Network & link analysis
Before 2000
Graph theory
After 1736
Present
small-world
networks
scale-free
networks
1998
1999
node
influence
group
discovery
network
modeling
Fraud detection
Software engineering
Bibliometrics
2008
Present
Applications to large business

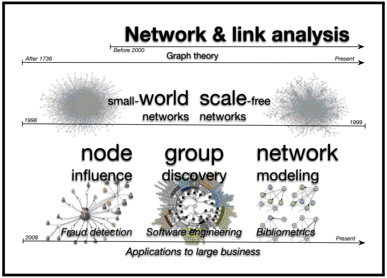
Applications to large business

Postdoc project 2014–2015:

Algorithms for network analysis in large company

Project outline/2:





Lovro Šubelj

University of Ljubljana, Faculty of Computer and Information Science

`http://lovro.lpt.fri.uni-lj.si/`

`lovro.subelj@fri.uni-lj.si`

R. Albert, H. Jeong, and A. L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the ACM International Conference on Web Science*, pages 45–54, Evanston, IL, USA, 2012.

A. L. Barabási. *Linked: The new science of networks*. Perseus, 2002.

A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the AAAI International Conference on Weblogs and Social Media*, pages 361–362, San Jose, CA, USA, 2009.

P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5): 1170–1182, 1987.

H. Burch and B. Cheswick. Network of autonomous systems of internet by internet mapping project.

A. Davis, B. B. Gardner, and M. R. Gardner. *Deep South*. Chicago University Press, Chicago, 1941.

W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, Cambridge, 2005.

C. I. Del Genio, T. Gross, and K. E. Bassler. All scale-free networks are sparse. *Phys. Rev. Lett.*, 107(17):178701, 2011.

P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in clobal social networks. *Science*, 301(5634):827–829, 2003.

M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Comput. Commun. Rev.*, 29(4):251–262, 1999.

L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

L. C. Freeman. Centrality in social networks: Conceptual clarification. *Soc. Networks*, 1(3): 215–239, 1979.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.

M. S. Granovetter. The strength of weak ties. *Am. J. Sociol.*, 78(6):1360–1380, 1973.

A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the Python in Science Conference*, pages 11–15, Pasadena, CA, USA, 2008.

D. Hansen, B. Shneiderman, and M. A. Smith. *Analyzing Social Media Metworks with NodeXL: Insights from a Connected World*. Morgan Kaufmann, Burlington, 2010.

R. Hobbs and M. Lombardi. *Mark Lombardi: Global Networks*. Independent Curators International, New York, 2003.

KDD. ar{X}iv citation network on high energy particle physics ({kdd}-{c}up 2003 dataset), 2003.

J. Kleinberg. Small-world phenomena and the dynamics of information. In *Proceedings of the International Conference in Advances in Neural Information Processing Systems*, 2001.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

P. L. Krapivsky and S. Redner. Network growth by copying. *Phys. Rev. E*, 71(3):036118, 2005.

P. J. Laurienti, K. E. Joyce, Q. K. Telesford, J. H. Burdette, and S. Hayasaka. Universal fractal scaling of self-organized networks. *Physica A*, 390(20):3608–3613, 2011.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):1–41, 2007.

A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Inf. Retr.*, 3(2):127–163, 2000.

M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: Patterns and a generator. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 524–532, New York, NY, USA, 2008.

S. Milgram. The small world problem. *Psychol. Today*, 1(1):60–67, 1967.

J. L. Moreno. *Who Shall Survive?* Beacon House, Beacon, 1934.

M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.

M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *P. Natl. Acad. Sci. USA*, 104(23):9564, 2007.

M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, 2003.

J. O'Madadhain, D. Fisher, S. White, P. Smyth, and Y.-B. Boey. Analysis and visualization of network data using JUNG. *J. Stat. Softw.*, 10(2):1–35, 2005.

L. Page. Method for node ranking in a linked database, 2001.

J. D. Pelletier. Self-organization and scaling relationships of evolving river networks. *Journal of Geophysical Research*, 104(4):7359–7375, 1999.

D. J. d. S. Price. Networks of scientific papers. *Science*, 149:510–515, 1965.

U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76(3):036106, 2007.

J. Scott. *Social Network Analysis: A Handbook*. Sage, London, 2 edition, 2000.

M. V. Simkin and V. P. Roychowdhury. Read before you cite! *Compl. Syst.*, 14:269–274, 2003.

S. Sinha. Few and far between. *Physics*, 4:81, 2011.

SNAP. Stanford network analysis project (SNAP), 2013.

L. Šubelj. *Odkrivanje goljufij na osnovi analize socialnih mrež*. BSc thesis, University of Ljubljana, Ljubljana, Slovenia, 2008.

L. Šubelj. *Odkrivanje skupin vozlišč v velikih realnih omrežjih na osnovi izmenjave oznak*. PhD thesis, University of Ljubljana, Ljubljana, Slovenia, 2013.

L. Šubelj and M. Bajec. Robust network community detection using balanced propagation. *Eur. Phys. J. B*, 81(3):353–362, 2011a.

L. Šubelj and M. Bajec. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. *Phys. Rev. E*, 83(3):036103, 2011b.

L. Šubelj and M. Bajec. Community structure of complex software systems: Analysis and applications. *Physica A*, 390(16):2968–2975, 2011c.

L. Šubelj and M. Bajec. Software systems through complex networks science: Review, analysis and applications. In *Proceedings of the KDD Workshop on Software Mining*, pages 9–16, Beijing, China, 2012a.

L. Šubelj and M. Bajec. Ubiquitousness of link-density and link-pattern communities in real-world networks. *Eur. Phys. J. B*, 85(1):32, 2012b.

L. Šubelj and M. Bajec. Model of complex networks based on citation dynamics. In *Proceedings of the WWW Workshop on Large Scale Network Analysis*, pages 527–530, Rio de Janeiro, Brazil, 2013.

L. Šubelj and M. Bajec. Group detection in complex networks: An algorithm and comparison of the state of the art. *Physica A*, 397:144–156, 2014.

L. Šubelj, Š. Furlan, and M. Bajec. Odkrivanje goljufij na osnovi analize socialnih omrežij. In *Zbornik Konference Dnevi Slovenske Informatike*, page 10, Portorož, Slovenia, 2009.

L. Šubelj, Š. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Syst. Appl.*, 38(1):1039–1052, 2011.

L. Šubelj, N. Blagus, and M. Bajec. Group extraction for real-world networks: The case of communities, modules, and hubs and spokes. In *Proceedings of the International Conference on Network Science*, pages 152–153, Copenhagen, Denmark, 2013a.

L. Šubelj, S. Žitnik, N. Blagus, and M. Bajec. Node mixing and group structure of complex software networks. *sub. to Adv. Complex Syst.*, page 23, 2013b.

S. Valverde, R. F. Cancho, and R. V. Solé. Scale-free networks from optimal design. *Europhys. Lett.*, 60(4):512–517, 2002.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684):440–442, 1998.

J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode caenorhabditis elegans. *Phil. Trans. R. Soc. Lond. B*, 314(1165): 1–340, 1986.

W. W. Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33(4):452–473, 1977.

Y. Zhao, E. Levina, and J. Zhu. Community extraction for social networks. *P. Natl. Acad. Sci. USA*, 108(18):7321–7326, 2011.