A Graph Degeneracy-based Approach to Keyword Extraction

Antoine J.-P. Tixier[1], Fragkiskos D. Malliaros[1,2], Michalis Vazirgiannis[1]

[1]Computer Science Laboratory, École Polytechnique, Palaiseau, France
[2]Department of Computer Science and Engineering, UC San Diego, La Jolla, CA, USA

Motivation

Graph-degeneracy is better than PageRank for keyword extraction [Rousseau & Vazirgiannis 2015], but:

- retaining only the main core is **suboptimal**: one cannot expect all the keywords to live in the top level of the hierarchy

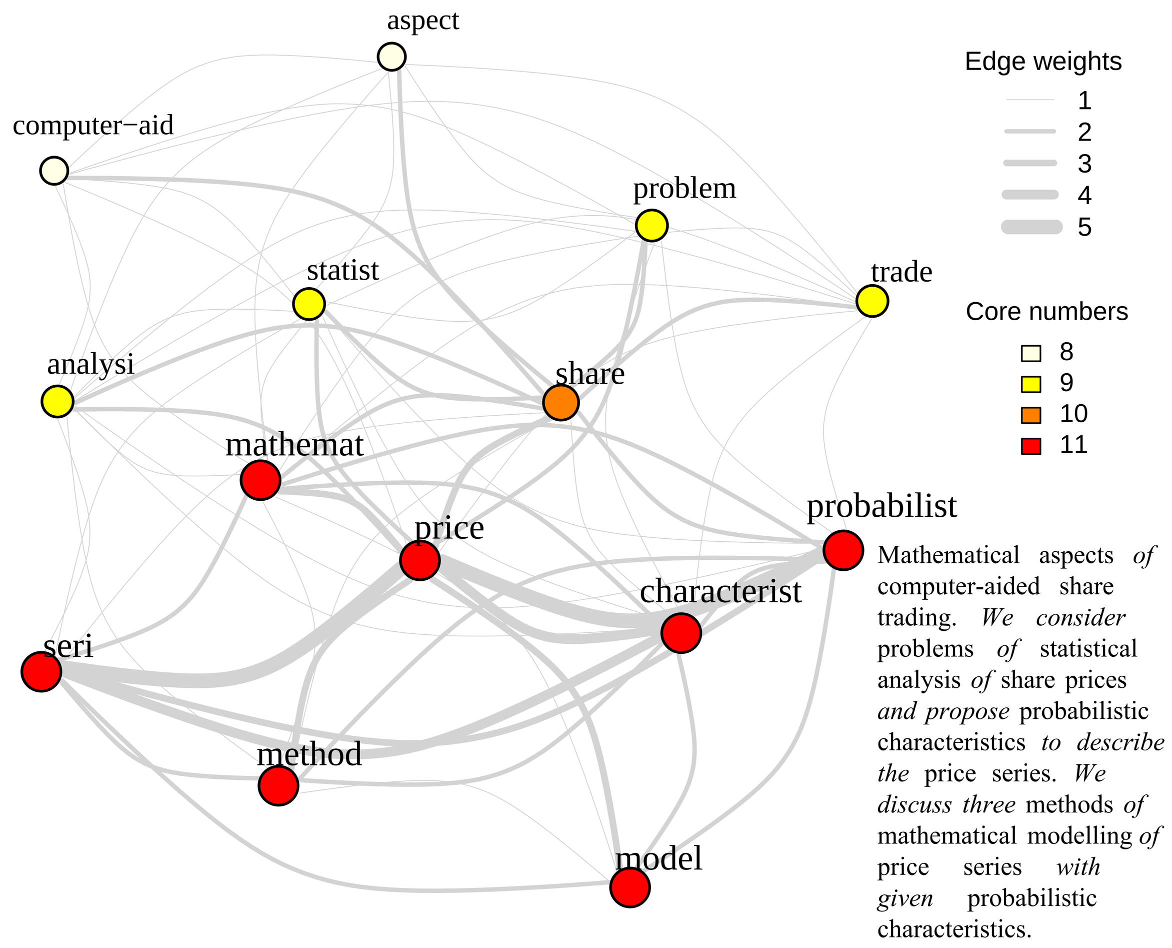
  ↪ ***how to automatically select the best hierarchy level?***

  - ***dens***: go down the hierarchy until a drop in density is observed
  - ***inf***: go down the hierarchy as long as the shells ↗ in size

- working with **subgraphs** lacks flexibility

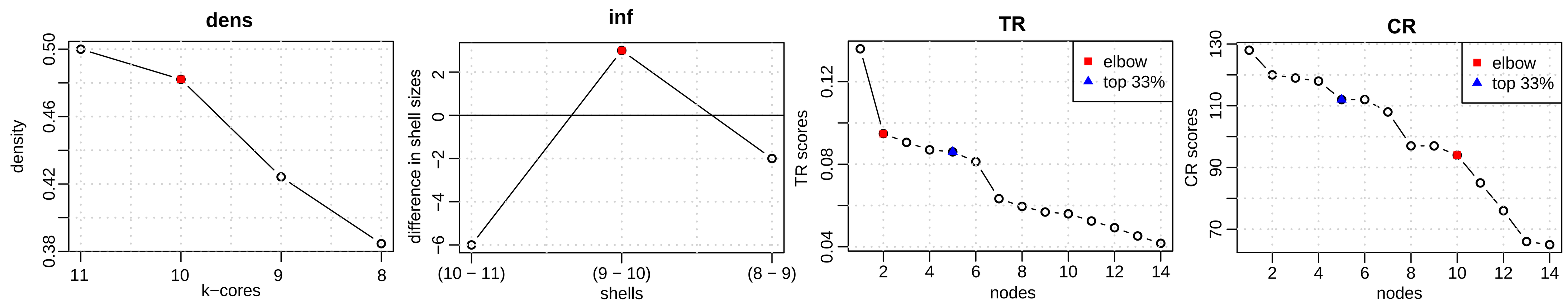
  ↪ ***how to rank nodes individually while retaining the valuable cohesiveness information captured by degeneracy?***

  - ***CoreRank (CR)***: (1) assign to each node the sum of the core or truss numbers of its neighbors, (2) select the elbow in the scores curve (***CRE***) or retain the top p% nodes (***CRP***)



Core numbers				TR scores				CR scores			
	P	R	F1		P	R	F1		P	R	F1
MAIN	0.86	0.55	0.67	ELB	1	0.18	0.31	ELB	0.90	0.82	0.86
INF	**0.83**	**0.91**	**0.87**	PER	1	0.45	0.63	PER	1	0.45	0.63
DENS	0.88	0.64	0.74								

Core numbers		TR scores		CR scores	
mathemat	11	**price**	.1359	**mathemat**	128
price	11	**share** *ELB*	.0948	**price**	120
probabilist	11	**probabilist**	.0906	**analysi**	119
characterist	11	**characterist**	.0870	**share**	118
seri	11	**seri** *PER*	.0860	**probabilist** *PER*	112
method	11	**mathemat**	.0812	**characterist**	112
model *MAIN*	11	**analysi**	.0633	**statist**	108
share *DENS*	10	**statist**	.0595	**trade**	97
trade	9	method	.0569	problem	97
problem	9	problem	.0560	**seri** *ELB*	94
statist	9	**trade**	.0525	method	85
analysi *INF*	9	**model**	.0493	**computer-aid**	76
aspect	8	**computer-aid**	.0453	**model**	66
computer-aid	8	aspect	.0417	aspect	65

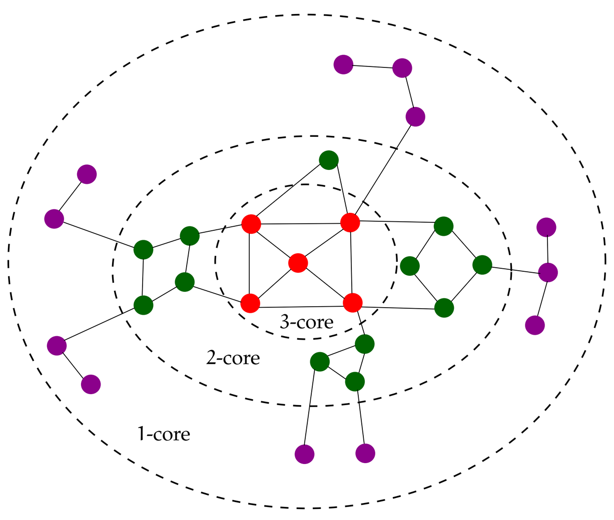


$W = 8$. Weighted, undirected edges. weighted k-core. Nouns and adjectives as nodes.

Graph degeneracy

k-CORE DECOMPOSITION

- the k-core of $G = (V, E)$ is a maximal connected subgraph of G in which every **vertex** v has at least degree k [Seidman 1983]
- v has **core number** k if it belongs to the k-core but not to the $(k+1)$-core
- the k-core decomposition of G is the set of all its cores from $k = 0$ (G itself) to $k = k_{max}$ (its main core)
- **complexity**: $O(n+m)$ resp. $O(m \log(n))$ in time in the (un)weighted cases, $O(n)$ in space [Batagelj & Zaveršnik 2002]

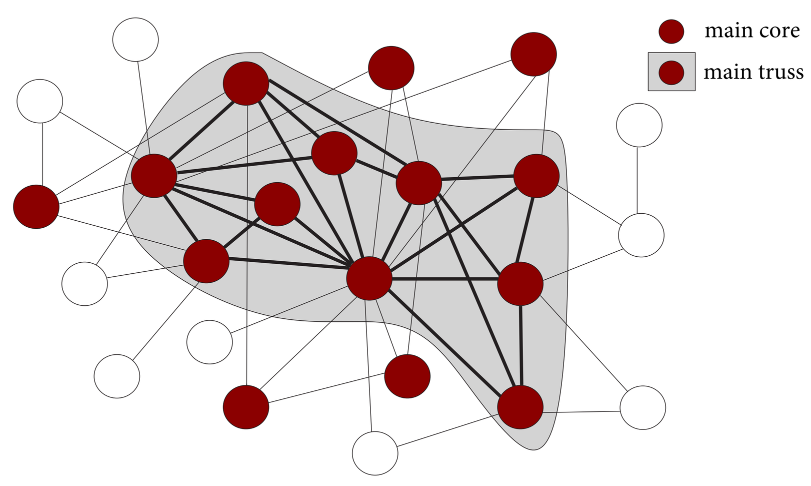


k-core decomposition

- hierarchy of nested subgraphs whose cohesiveness and size respectively ↗ and ↘ with k
- nodes with high core numbers are not only **central** but also form **cohesive subgraphs** with other central nodes

K-TRUSS DECOMPOSITION

- the K-truss of $G = (V, E)$ is its largest subgraph where every **edge** e belongs to at least $K - 2$ triangles [Cohen 2008]
- e has **truss number** K if it belongs to the K-truss but not to the $(K+1)$-truss
- the **truss number** of v is the maximum truss number of its adjacent edges
- the K-truss decomposition of G is the set of all its K-trusses from 2 (G) to K_{max}
- **complexity**: $O(m^{1.5})$ in time and $O(m+n)$ in space [Wang & Cheng 2012]



k-core versus K-truss

- compared to k-core, K-truss imposes constraints not only on the number of **direct links** but also on the number of **common neighbors**
- the K-trusses can be viewed as *cores* of the k-cores that filter out less cohesive elements [Wang & Cheng 2012]

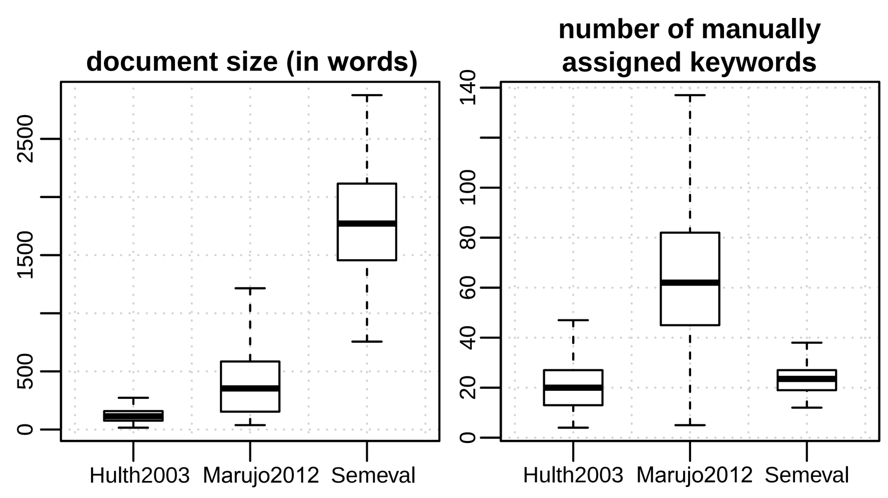
Degeneracy and Spreading Influence

- in social networks, the **best spreaders** are not the highly connected individuals, but those located at the **core of the network** [Kitsak 2010]
- the **truss number** is an even better indicator of spreading influence than the **core number** [Malliaros et al. 2016]

the spreading influence of a node is related to its **structural position** within the graph (*density* and *cohesiveness*) rather than to its **prestige** (*random walk*-based degree) ⇒ **influential words** should make better keywords

Datasets

- **Hulth2003**: 500 abstracts from the Inspec physics & engineering database
- **Marujo2012**: 450 web news stories covering 10 different topics
- **Semeval**: 100 scientific papers from the ACM



Results

For each data set, we retained the degeneracy technique and window size giving the absolute best performance

- our methods outperform all baselines by a wide margin
- drastic improvement in recall, for a comparatively lower loss in precision
- K-truss needs greater window sizes to perform well (more triangles)
- on long documents (Semeval), the lack of flexibility of subgraph-based approaches (***dens*** and ***inf***) is a handicap. Working at the node level (***CRP***) is better

	precision	recall	F1-score
dens	**48.79**	**72.78**	**56.09***
inf	48.96	72.19	55.98*
CRP	61.53	38.73	45.75
CRE	65.33	37.90	44.11
main†	51.95	54.99	50.49
TRP†	65.43	41.37	48.79
TRE†	71.34	36.44	45.77

Hulth2003, K-truss, $W = 11$. *stat. sign. ($p < 0.001$) w.r.t. all baselines†

	precision	recall	F1-score
dens	**47.62**	**71.46**	**52.94***
inf	53.88	57.54	49.10*
CRP	54.88	36.01	40.75
CRE	63.17	25.77	34.41
main†	64.05	34.02	36.44
TRP†	55.96	36.48	41.44
TRE†	65.50	21.32	30.68

Marujo2012, k-core, $W = 13$. *stat. sign. ($p < 0.001$) w.r.t. all baselines†

	precision	recall	F1-score
dens	8.44	79.45	15.06
inf	17.70	65.53	26.68
CRP	**49.67**	**32.88**	**38.98***
CRE	25.82	58.80	34.86
main†	25.73	49.61	32.83
TRP†	47.93	31.74	37.64
TRE†	33.87	46.08	37.55

Semeval, K-truss, $W = 20$. *stat. sign. ($p < 0.001$) w.r.t. *main*

{anti5662, fmalliaros, mvazirg}@lix.polytechnique.fr