ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

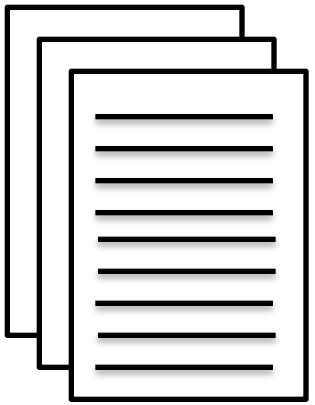
"Graph-Based Text Mining - I"

Michalis Vazirgiannis

DATA SCIENCE AND MINING GROUP,
LIX, École Polytechnique
http://www.lix.polytechnique.fr/dascim/

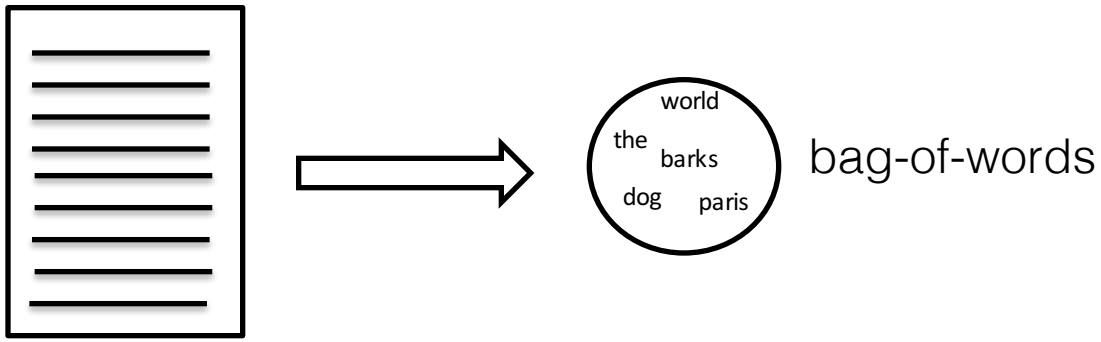
November, 2018

Text Mining – Terminology

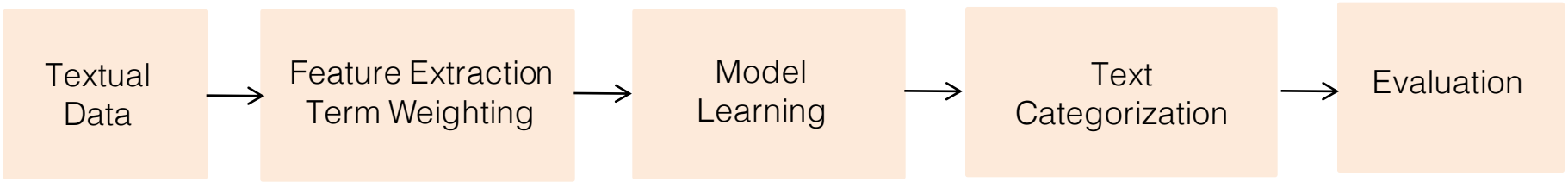


- Text mining on a collection of documents:
  - The collection is the data set
  - The documents are the data points

- Since text is unstructured, a document is usually converted in a common representation



Example: Text Categorization



Applications:

- Opinion mining (sentiment analysis)
- Email spam classification
- Web-pages classification
- …

Bag-of-Words (BoW) - Issues

Example text

information retrieval is the activity of obtaining

information resources relevant to an information need

from a collection of information resources

Bag of words: [(activity,1), (collection,1)
(information,4), (relevant,1),
(resources, 2), (retrieval, 1), …]

- Term independence assumption
- Term frequency weighting

Assumptions made by the BoW model

BM25 Ranking Function

- Ranking function assuming bag-of-words document representation

$$score(d,q) = \sum_{t \in d \cap q} idf_t \times \frac{tf_{t,d} \cdot (k_1 + 1)}{tf_{t,d} + k_1 \cdot \left(1 - b + b \frac{len_d}{avglen}\right)}$$

  - len_d is the length of document d
  - $avglen$ is the average document length in the collection

- Score depends only on query terms

- Values of parameters k_1 and b depend on collection/task
  - k_1 controls term frequency saturation
  - b controls length normalization
  - Default values: $k_1 = 1.2$ and $b = 0.75$

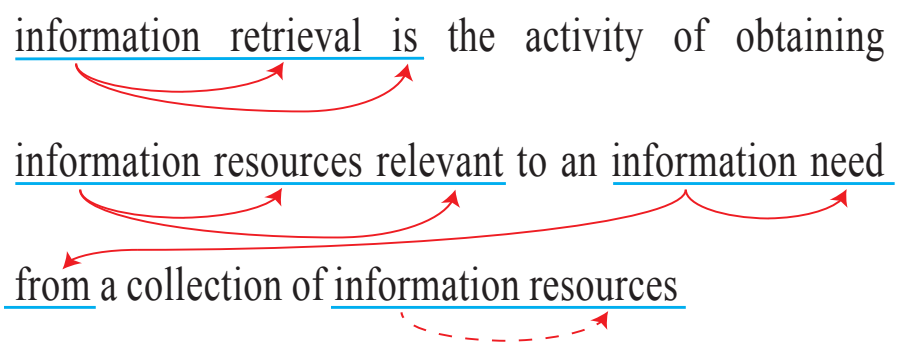
Graph-based Document Representation

- Challenge the term independence and term frequency weighting assumptions taking into account word dependence, order and distance

- Employ a graph-based document representation capturing the above

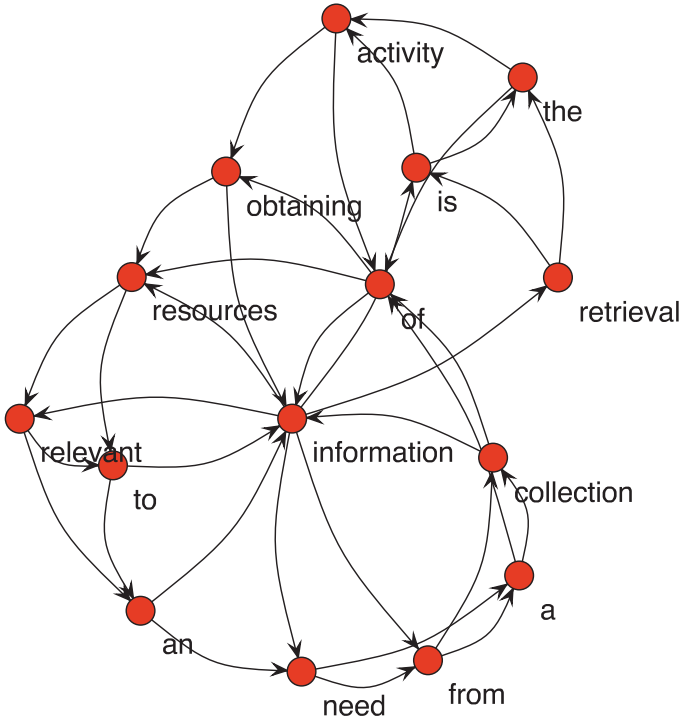
- Graphs have been successfully used in IR to encompass relations and propose meaningful weights (e.g., PageRank)

Graph-based Document Representation - Example



Idea: Replace term frequency with node centrality

Captures: frequency, order and distance



Goal of the Tutorial and Outline

Goal: offer a comprehensive presentation of recent methods that rely on graph-based text representations to deal with various tasks in NLP and IR

- Part I. Graph-theoretic concepts and graph-based text representation
- Part II. Information retrieval
- Part III. Keyword extraction and text summarization
- Part IV. Text categorization

Tutorial Outline

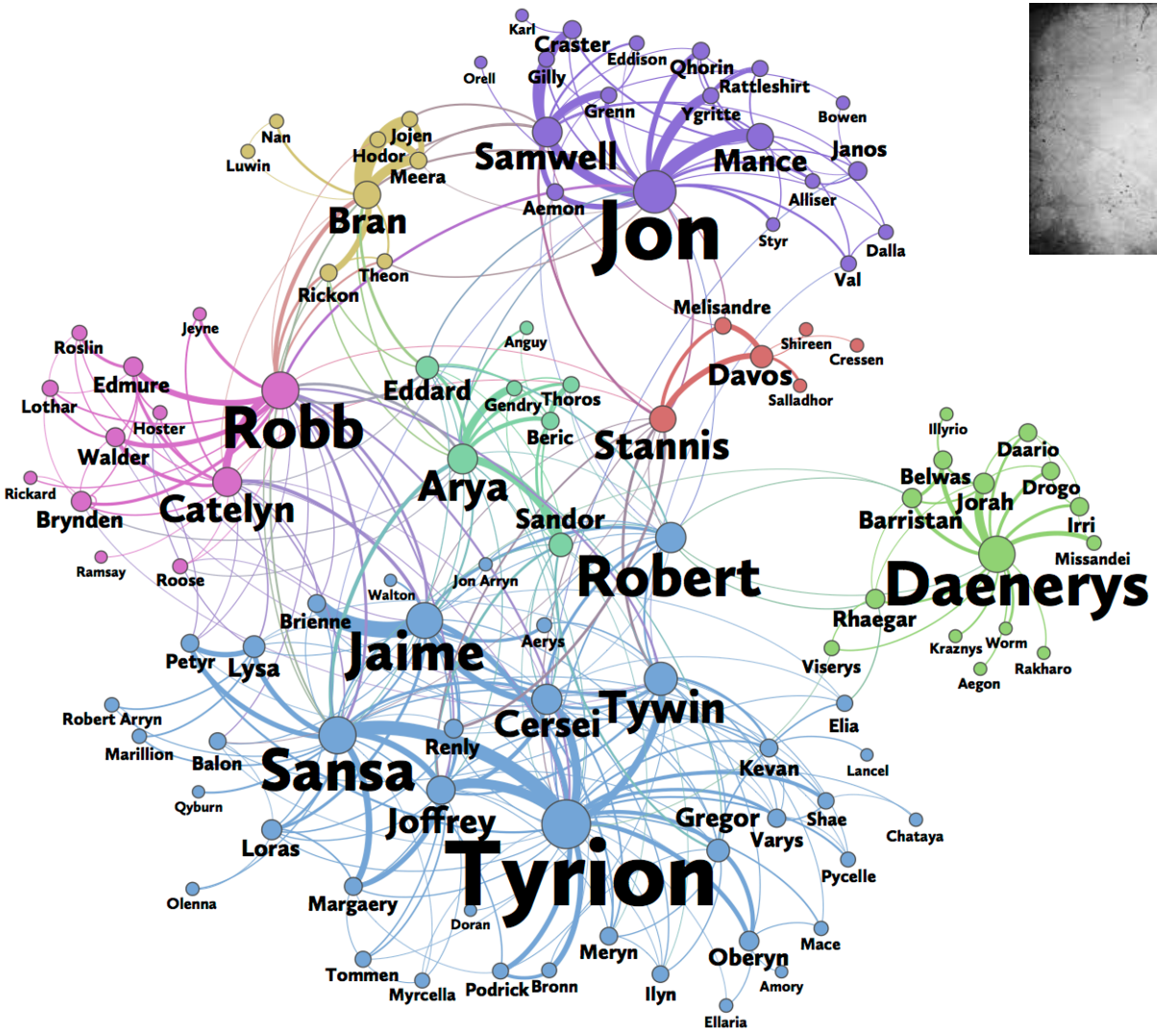
- Part I. Some useful concepts from Graph-theory
- Part II. Graph of Words - Information retrieval
- Part III. Graph based Event detection

Centrality criteria

Centrality in Networks (1/2)

- Determine the relative importance of a node in the network
  - Applications in Social Network Analysis, the Internet, Epidemiology, Urban informatics, …
- What do we mean by **centrality**?
  - A central node is more important or powerful …
  - Or, more influential ...
  - Or, is more critical due to its location in the graph
- Also, very closely related to the problem of **ranking** in the context of **Web search**
  - Each webpage can be considered as a 'user'
  - Each hyperlink is an endorsement relationship
  - **Centrality measures provide a query independent link-based score of importance of a web page**

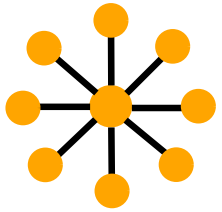
Centrality in Networks (2/2)



Source: https://www.macalester.edu/~abeverid/thrones.html

Types of Centrality

- **Starting point:** the central node of a **star** is the most important
- Why?
  - The node with the **highest degree**
  - The node that is closest to the rest nodes (e.g., has the smallest average distance to other nodes)
  - The node through which all shortest paths pass
  - The node that maximizes the dominant eigenvector (the one that corresponds to the largest eigenvalue) of the adjacency matrix
  - The node with highest probability in the stationary distribution of a random walk on the graph



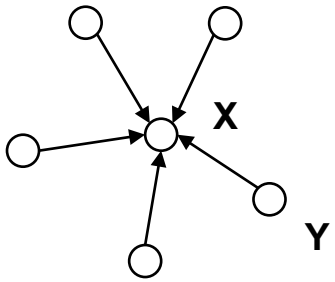
Various competing views of centrality

Measures of Centrality

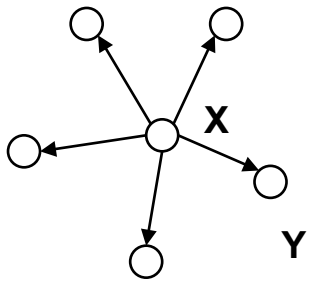
- This observation leads to the following classes of indices of centrality:
  - Measures based on **distances** (e.g., degree, closeness)
  - Measures based on **paths** (e.g., betweenness, Katz's index)
  - **Spectral** measures (eigenvector, PageRank, HITS, SALSA, random walks with restarts)
  - Measure based on **groups of nodes** (e.g., cliques, cores)
    - Related to the "clustering" structure
    - More on that in another lecture

A First Example

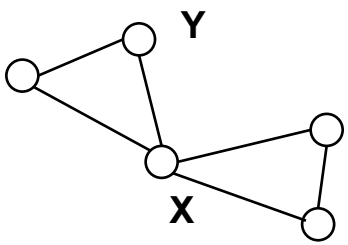
In each of the following networks, **X** has higher centrality than **Y** according to a particular measure



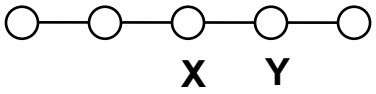
in-degree



out-degree



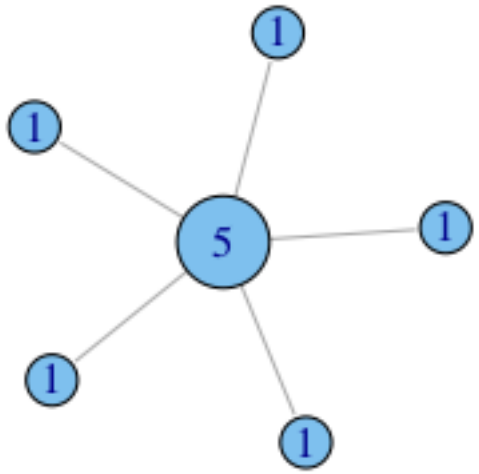
betweenness



closeness

Degree Centrality (1/2)

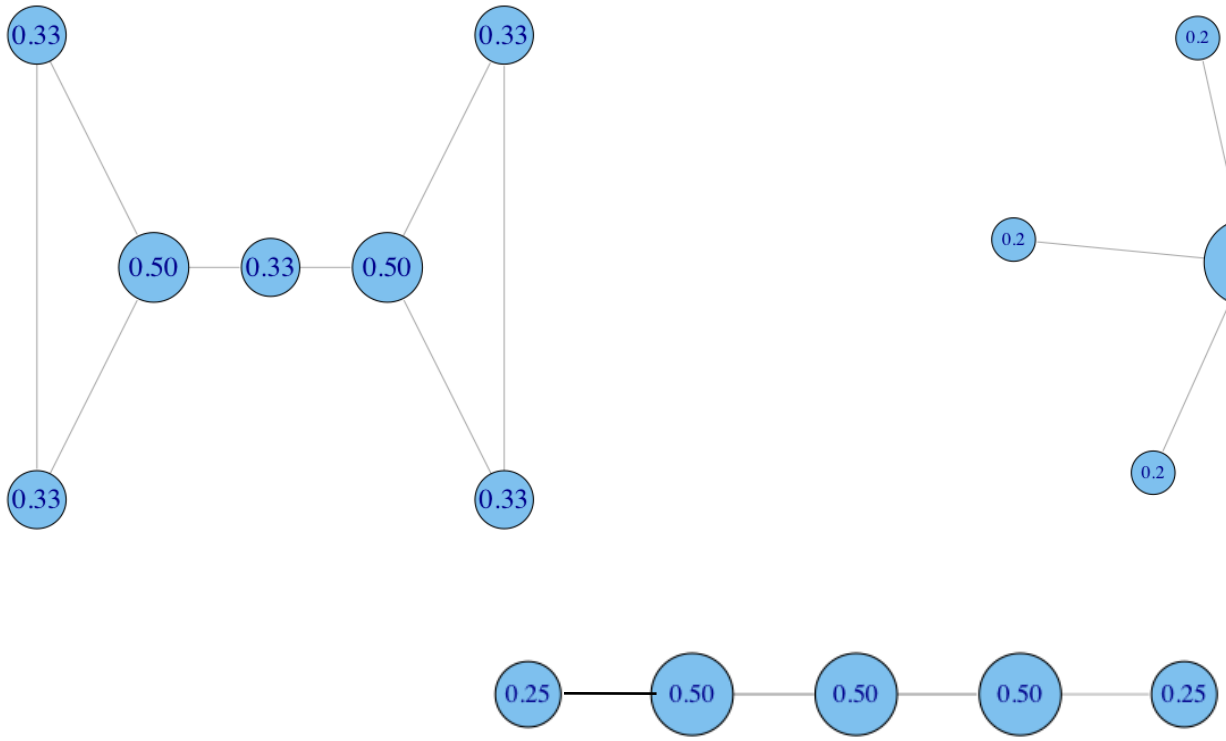
- **Idea:** A central node is one with many connections



- $C_d(i) = k(i)$, where $k(i)$ is the degree of node i

Degree Centrality (2/2)

- **Idea:** A central node is one with many connections



- Normalized degree centrality: divide by the max possible degree (***n-1***)

Closeness Centrality

- **Motivation:** it measures the ability to quickly access or pass information through the graph

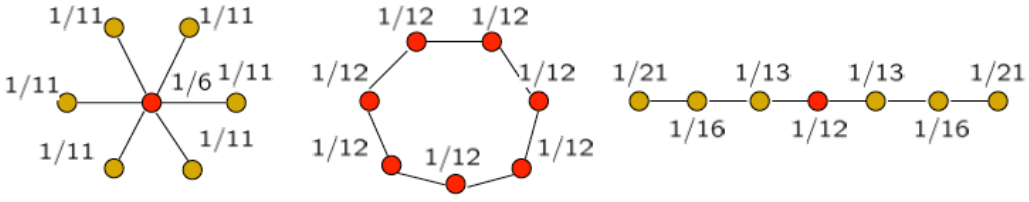
$$C_{Cl}(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}$$

Values in the range [0,1]

Mean distance from a node to other nodes

d(i, j) is the length of the shortest path between ***i*** and ***j*** (**geodesic distance**)

- The **closeness** of a node is defined as the **inverse** of the sum of the shortest path (SP) distances between the node and all other nodes in the graph



Be close to everybody else
(e.g., influence on other nodes)

Why inverse the distance?

- Nodes with **low mean distance** should get **high score**

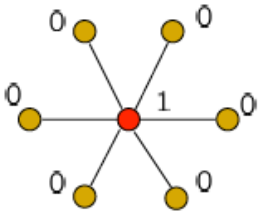
[Mateos, '17]

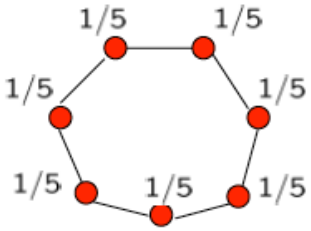
Betweenness Centrality

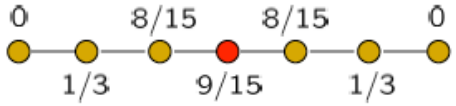
- **Motivation:** a node is important if it lies in many shortest paths

$$C_{bt}(i) = \sum_{s \neq i \neq t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)}$$

- ***σ(s, t)*** is the total number of shortest paths from ***s*** to ***t***
- ***σ(s, t\|v)*** is the number of shortest paths from ***s*** to ***t*** that pass through ***i***







Essential nodes in passing information through the network

Oftentimes it is normalized: $\frac{C_{bt}(i)}{\binom{n-1}{2}}$

Core decomposition in networks

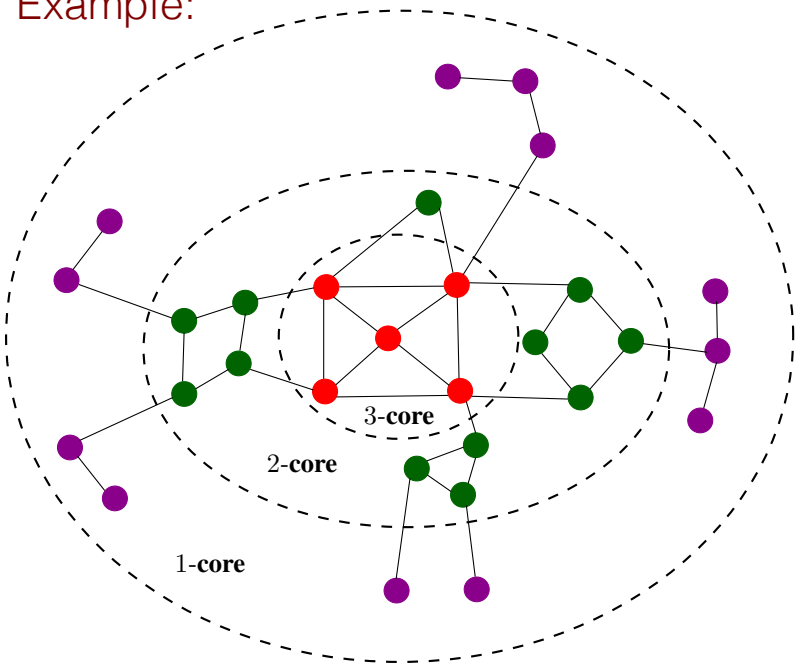
Core Decomposition

- Tool to analyze the structure of real networks
  - Quantify community and clustering structure
- Hierarchical representation of a graph into nested subgraphs of increased connectivity and coherence properties
- **Basic idea:**
  - Set a threshold on the node degree, say **k**
  - Nodes that do not satisfy the threshold are removed from the graph
- Extensions to other node properties (e.g., triangles)
- Plethora of applications
  - Dense subgraph discovery and community detection
  - Evaluation of collaboration in social networks
  - Identification of influential spreaders in social networks
  - **Text analytics**

k-Core Decomposition

- Degeneracy for an undirected graph **G**
  - Also known as the **k**-core number
  - The **k**-core of **G** is the largest subgraph in which every vertex has degree at least **k** within the subgraph

Example:



Core number $c_i = 1$

Core number $c_i = 2$

Core number $c_i = 3$

Graph Degeneracy $\delta^*(G) = 3$

$G_0 = G$

$G_1 = 1$-core of G

$G_2 = 2$-core of G

$G_3 = 3$-core of G

$G_0 \supseteq G_1 \supseteq G_2 \supseteq G_3$

Important property:

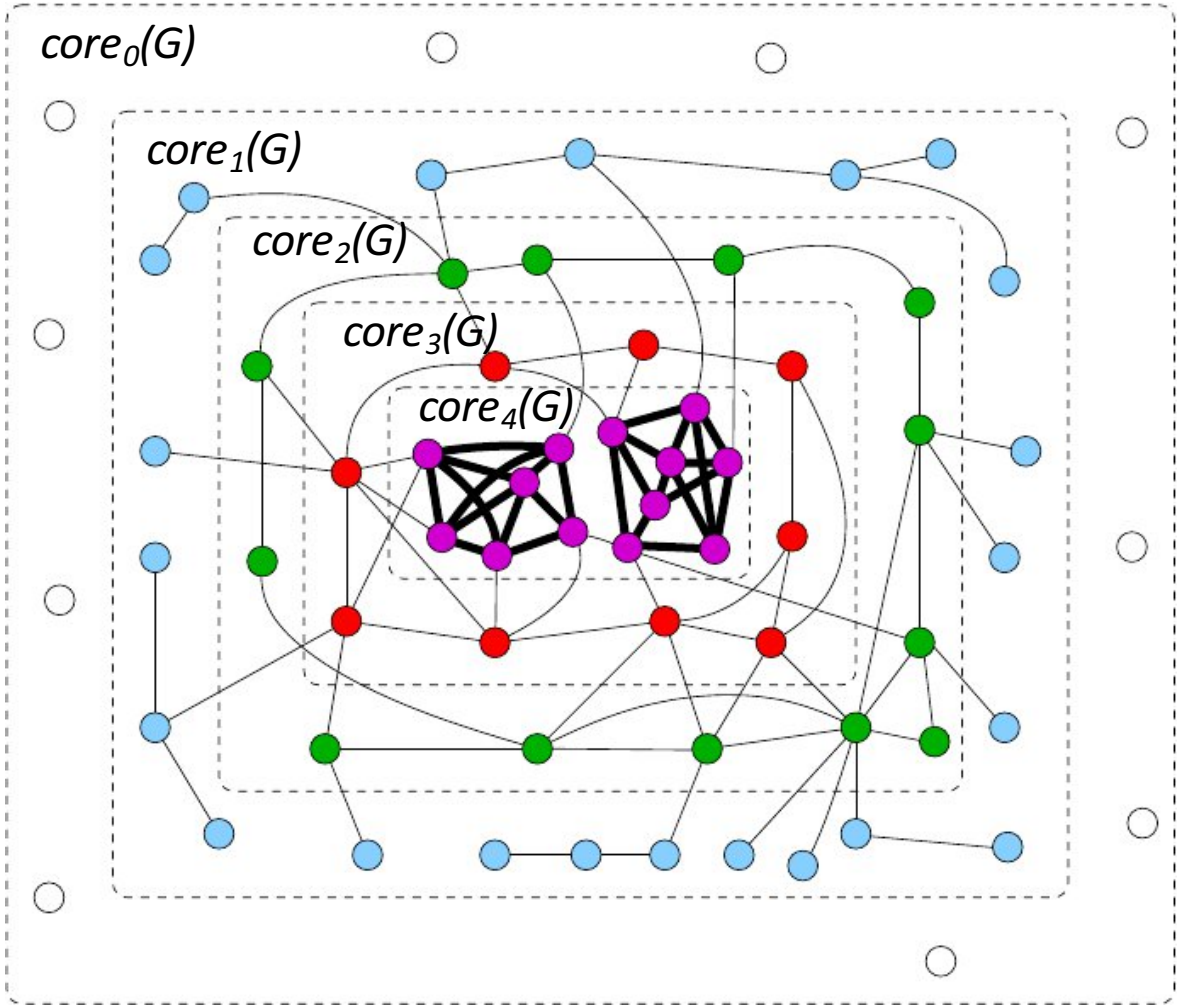
- Fast and easy to compute
- Linear to the size of the graph
- Scalable to large scale graphs

Note:

The degeneracy and the size of the k-core provide a good indication of the cohesiveness of the graph

Also known as graph degeneracy

Another Example



Algorithm for k-Core Decomposition

Algorithm k-core(G, k)

Input: An undirected graph G and positive integer k

Output: k-core(G)

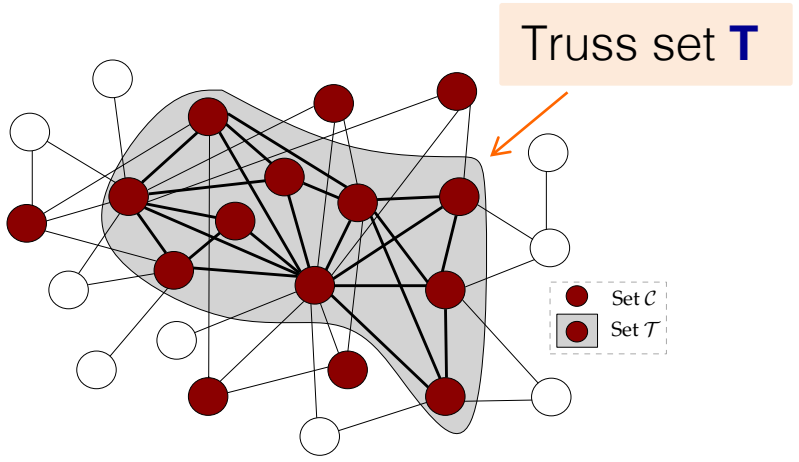
1. let F := G
2. while there is a node x in F such that $\deg_F(x)<k$
   delete node x from F
3. return F

- Many efficient algorithms have been proposed for the computation
  - Time complexity: **O(m)**

[Batagelj and Zaversnik, '03]

K-truss Decomposition (Triangles)

- K-truss decomposition [Cohen '08], [Wang and Cheng '12]
  - **Triangle-based** extension of the **k**-core decomposition
  - Each edge of the K-truss subgraph participates in at least **K-2** triangles
    - Informally, the "core" of the maximal **k**-core subgraph
    - Subgraph of higher coherence compared to the **k**-core



Graph-based text representations

Graph Semantics

- Let **G = (V, E)** be the graph that corresponds to a document **d**
- The nodes can correspond to:
  - Paragraphs
  - Sentences
  - Phrases
  - Words [Main focus of the tutorial]
  - Syllables
- The edges of the graph can capture various types of relationships between two nodes:
  - Co-occurrence within a window over the text [Main focus of the tutorial]
  - Syntactic relationship
  - Semantic relationship

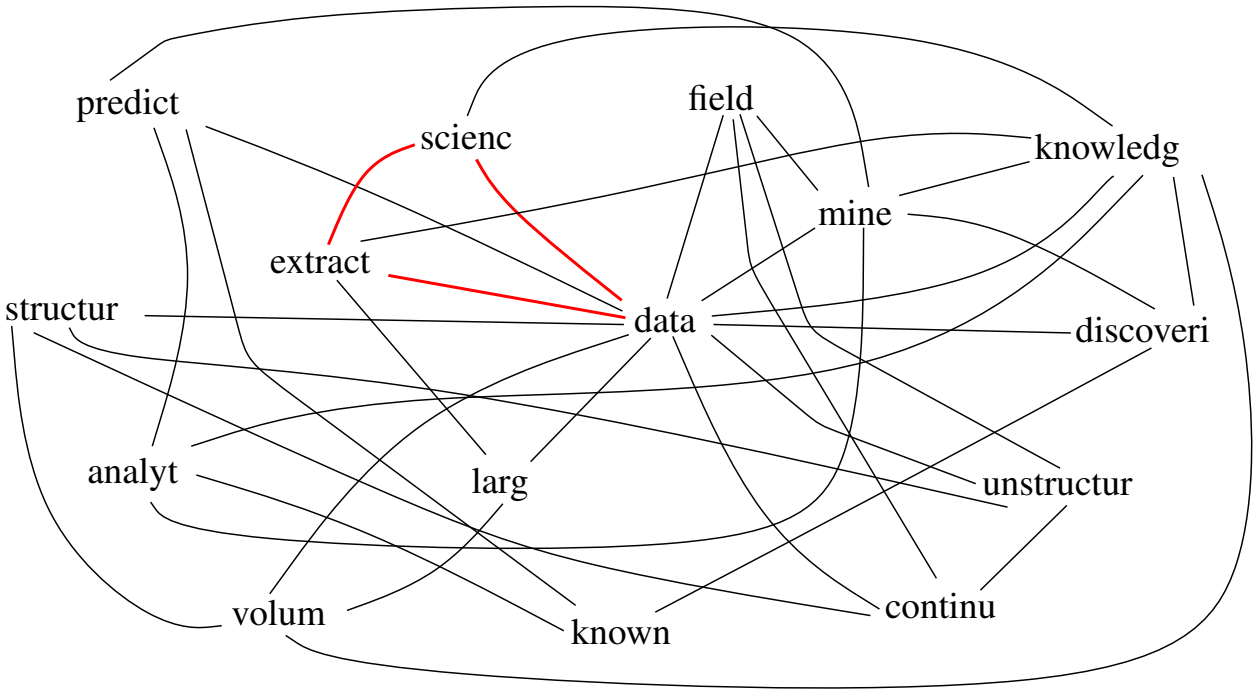
Graph-of-Words (GoW) Model

- Each document **d** ∈ **D** is represented by a graph $G_d = (V_d, E_d)$, where the nodes correspond to the terms **t** of the document and the edges capture co-occurrence relationships between terms within a fixed-size sliding window of size **w**
- Directed vs. undirected graph
  - Directed graphs are able to preserve the actual flow of a text
  - In undirected graphs, an edge captures co-occurrence of two terms whatever the respective order between them is
- Weighted vs. unweighted graph
  - The higher the number of co-occurrences of two terms in the document, the higher the weight of the corresponding edge
- Size **w** of the sliding window
  - Add edges between the terms of the document that co-occur within a sliding window of size **w**
  - Larger window sizes produce graphs that are relatively dense

[Mihalcea and Tarau, EMNLP '04], [Blanco and Lioma, Inf. Retr. '12], [Rousseau and Vazirgiannis, CIKM '13]

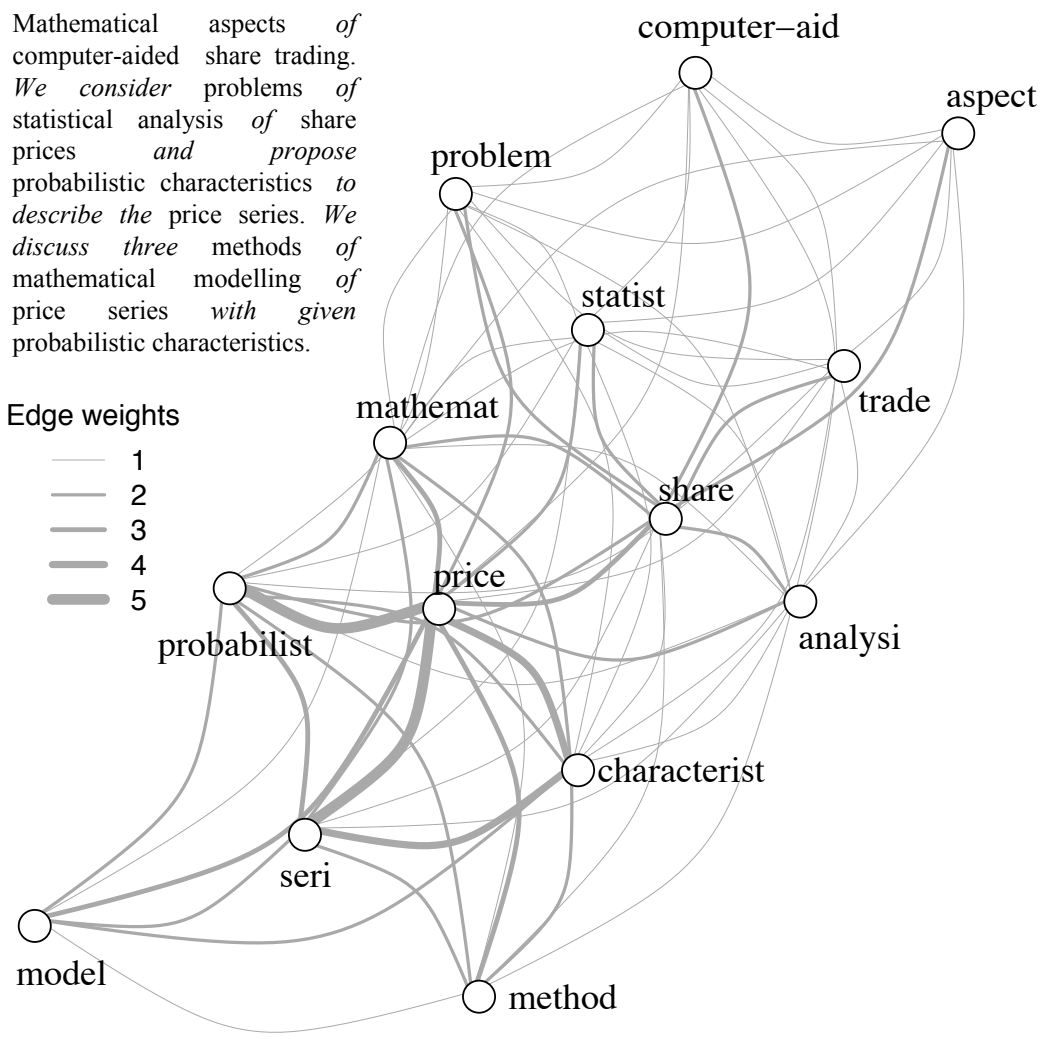
Example of Unweighted GoW

Data Science ~~is the~~ extraction of knowledge from large volumes of data that are structured or unstructured which is a continuation of the field of data mining and predictive analytics, also known as knowledge discovery and data mining.



w = 3
unweighted, undirected graph

Example of Weighed Undirected GoW



Tutorial Outline

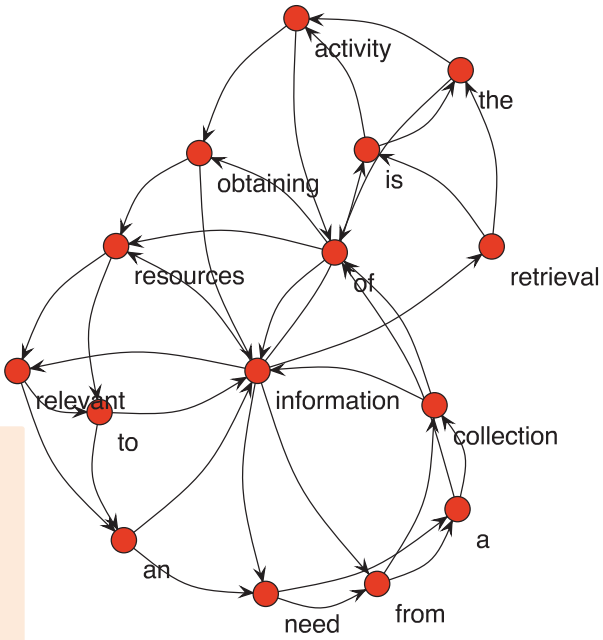
- Part I. Some useful concepts from Graph-theory
- Part II. Graph of Words - Information retrieval
- Part III. Graph based Event detection

In-degree based TW

- The weight of a term in a document is its in-degree in the graph-of-words
- It represents the number of distinct contexts of occurrence
- We store the document as a vector of weights in the direct index and similarly in the inverted index
- For example:

information	5
retrieval	1
is	2
the	2
activity	2
of	3
obtaining	2
resources	3
relevant	2
to	2
an	2
need	2
from	2
a	2
collection	2

Bag of words:
((activity,1), (collection,1), (information,4), (relevant,1), (resources, 2), (retrieval, 1)..)



TF-IDF and BM25

- Term **t**, document **d**, collection of size **N**, term frequency **tf(t,d)**, document frequency **df(t)**, document length **ldl**, average document length **avdl**, asymptotical marginal gain k_1 (1.2), slope parameter **b**

- TF-IDF [Singhal et al., TREC-7]

$$\text{TF-IDF}(t, d) = \text{TF}_{\text{pol}}\text{-IDF}(t, d) = \text{TF}_{\text{p}}\text{oTF}_{\text{l}}(t, d) \text{ x } \text{IDF}(t) = \left(\frac{1 + \log\left(1 + \log\left(tf(t,d)\right)\right)}{1 - b + b \times \frac{|d|}{avdl}} \right) \times \log\left(\frac{N+1}{df(t)} \right)$$

- BM25 [Lv and Zhai, CIKM '11]

$$\text{BM25}(t, d) = \left(\frac{(k_1 + 1) \times tf(t,d)}{k_1 \times \left(1 - b + b \times \frac{|d|}{avdl}\right) + tf(t,d)} \right) \times \log\left(\frac{N+1}{df(t)} \right)$$

TW-IDF

- Term **t**, document **d**, collection of size **N**, term weight **tw(t, d)**, document frequency **df(t)**, document length **ldl**, average document length **avdl**, asymptotical marginal gain $\mathbf{k_1}$ (1.2), slope parameter **b**

$$\text{TW-IDF(t, d)} = \left(\frac{tw(t,d)}{1 - b + b \times \frac{|d|}{avdl}} \right) \times \log\left(\frac{N+1}{df(t)} \right)$$

- In the bag-of-word representation, **tw** is usually defined as the term frequency or sometimes just the presence/absence of a term (binary **tf**)
- In the graph-of-word representation, **tw** is the **in-degree** of the vertex representing the term in the graph

[Rousseau and Vazirgiannis, CIKM '13]

Experimental Evaluation

- Datasets
- Platforms
- Evaluation
- Results

Datasets (1/2)

- **Disks 1 & 2 (TREC)**
  741,856 news articles from Wall Street Journal (1987-1992), Federal Register (1988-1989), Associated Press (1988-1989 and Information from the Computer Select disks (1989-1990)
- **Disks 4 & 5 (TREC, minus the Congressional Record)**
  528,155 news releases from Federal Register (1994), Financial Times (1991-1994), Foreign Broadcast Information Service (1996) and Los Angeles Times (1989-1990)
- **WT10G (TREC)**
  1,692,096 crawled pages from a snapshot of the Web in 1997
- **.GOV2 (TREC)**
  25,205,179 crawled Web pages from .gov sites in early 2004

Datasets (2/2)

Dataset / Statistic	Disks 1 & 2	Disks 4 & 5	WT10G	.GOV2
# of documents	741,856	528,155	1,692,096	25,205,179
# of unique terms	535,001	520,423	3,135,780	15,324,292
average # of terms (avdl)	237	272	398	645
average # of vertices	125	157	165	185
average # of edges	608	734	901	1,185

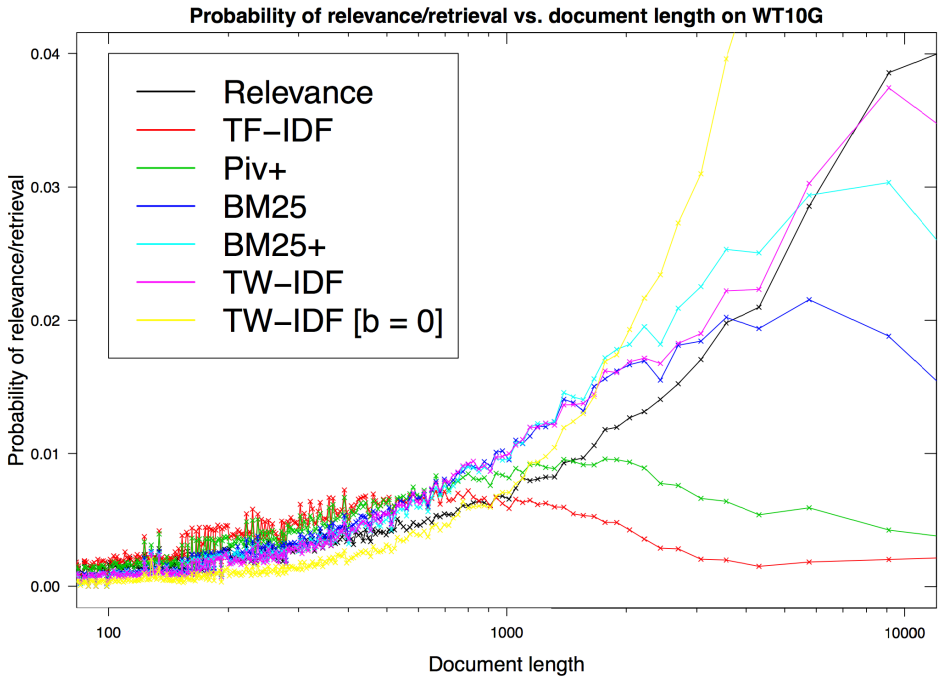
Table: Statistics on the four TREC datasets used; Disks 4&5 excludes the Congressional Record. The average values are computed per document.

Evaluation

- Mean Average Precision (MAP) and Precision at 10 (P@10)
  - Considering only the top-ranked 1000 documents for each run
- Statistical significance of improvement was assessed using the Student's paired t-test
  - R implementation (t.test {stats} package), trec_eval output as input
  - Two-sided p-values less than 0.05 and 0.01 to reject the null hypothesis
- Likelihood of relevance vs. likelihood of retrieval [Singhal et al., SIGIR '96]
- 4 baseline models: TF-IDF, BM25, Piv+ and BM25+
  - Tuned slope parameter *b* for pivoted document length normalization (2-fold cross-validation, odd vs. even topic ids, MAP maximization)
  - Default (1.0) lower-bounding gap [Lv and Zhai, CIKM '11]

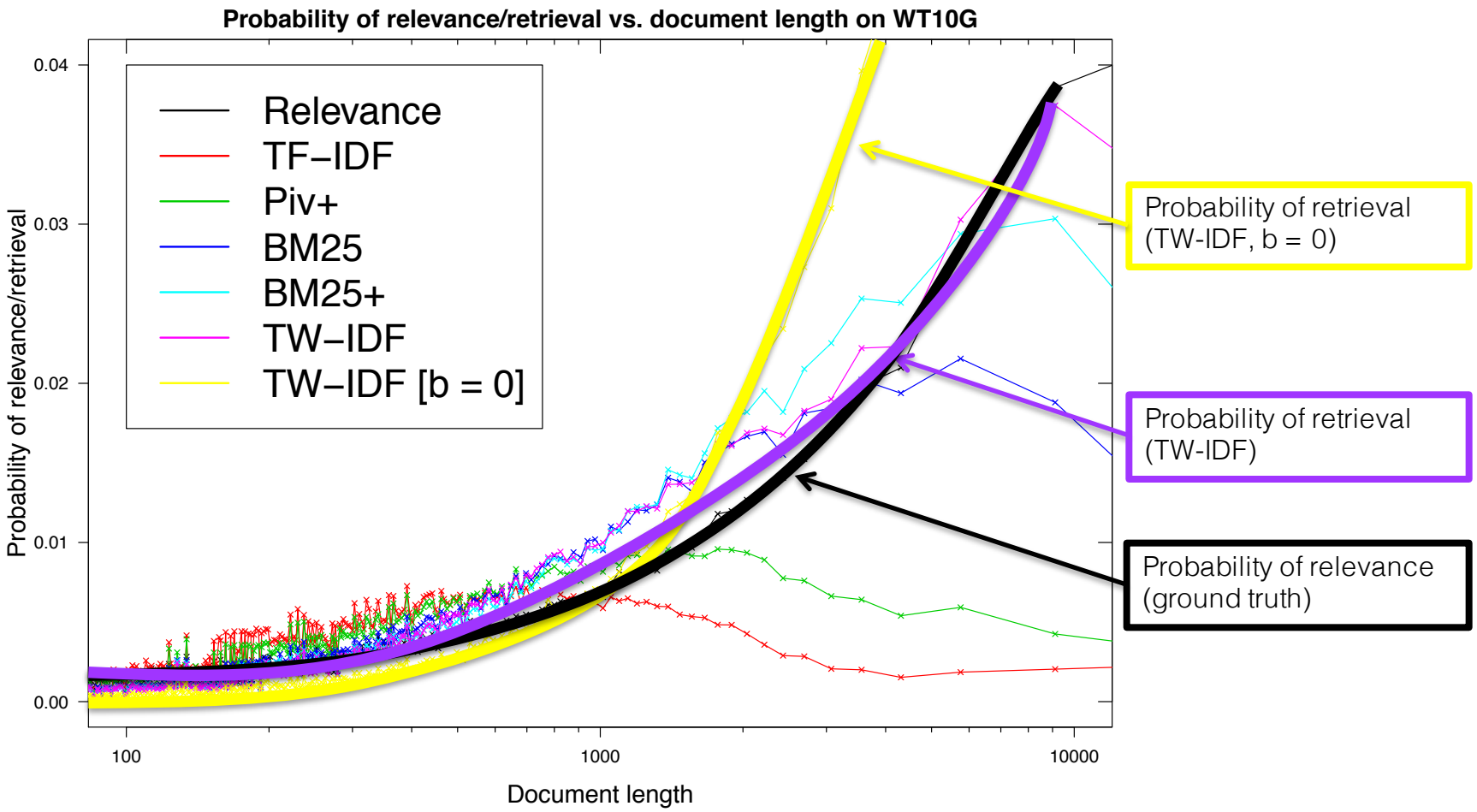
Graph-based Ad Hoc IR

- Evaluation in terms of:
  - Mean Average Precision
  - Precision@10
  - Probability of relevance vs. probability of retrieval



Model	b	TREC1-3 Ad Hoc		TREC 2004 Robust		TREC9-10 Web		TREC 2004-2006 Terabyte	
		MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
TF_{pol}	0.20	0.1471	0.3960	0.1797	0.3647	0.1260	0.1875	0.1853	0.4913
TF_{kop}	0.75	0.1346	0.3533	0.2045	0.3863	0.1702	0.2208	0.2527	0.5342
TW	none	0.1502	0.3662	0.1809	0.3273	0.1430	0.1979	0.2081	0.5021
TW_p	0.003	**0.1576****	**0.4040****	**0.2190****	**0.4133****	**0.1946****	**0.2479****	**0.2828****	**0.5407****
TF-IDF	0.20	0.1832	0.4107	0.2132	0.4064	0.1430	0.2271	0.2068	0.4973
BM25	0.75	0.1660	0.3700	0.2368	0.4161	0.1870	0.2479	0.2738	0.5383
TW-IDF	0.003	**0.1973****	**0.4148***	**0.2403****	**0.4180***	**0.2125****	**0.2917****	**0.3063****	**0.5633****

Likelihood of Relevance vs. Likelihood of Retrieval



Tutorial Outline

- Part I. Some useful concepts from Graph-theory
- Part II. Graph of Words - Information retrieval
- Part III. Graph based Event detection

Single Document Keyword Extraction

Keywords are used everywhere

- Looking up information on the Web (e.g., via a search engine bar)
- Finding similar posts on a blog (e.g., tag cloud)
- For ads matching (e.g., AdWords' keyword planner)
- For research paper indexing and retrieval (e.g., SpringerLink)
- For research paper reviewer assignment

Applications are numerous

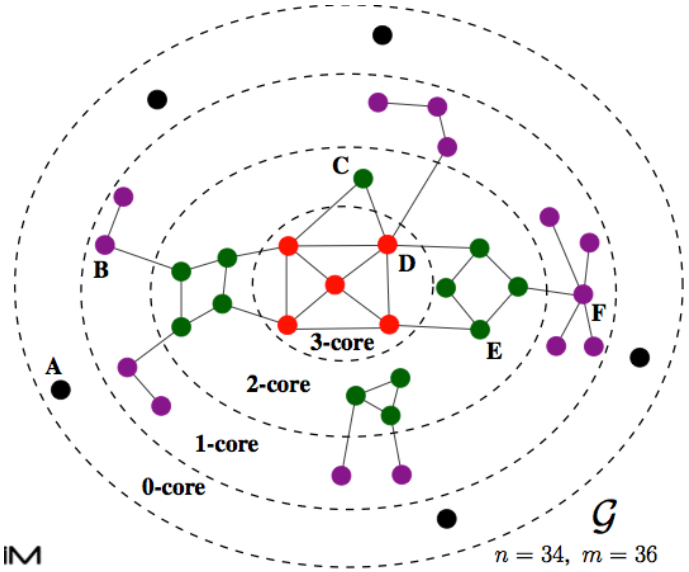
- **Summarization** (to get a gist of the content of a document)
- **Information filtering** (to select specific documents of interest)
- **Indexing** (to answer keyword-based queries)
- **Query expansion** (using additional keywords from top results)

Graph-based Keyword Extraction (1/2)

Existing graph-based keyword extractors:

- Assign a centrality based score to a node
- Top ranked ones will correspond to the most representative

- TextRank (PageRank) [Mihalcea and Tarau, EMNLP '04]
- HITS [Litvak and Last, MMIES '08]
- Node centrality (degree, betweenness, eigenvector) [Boudin, IJNLP '13]

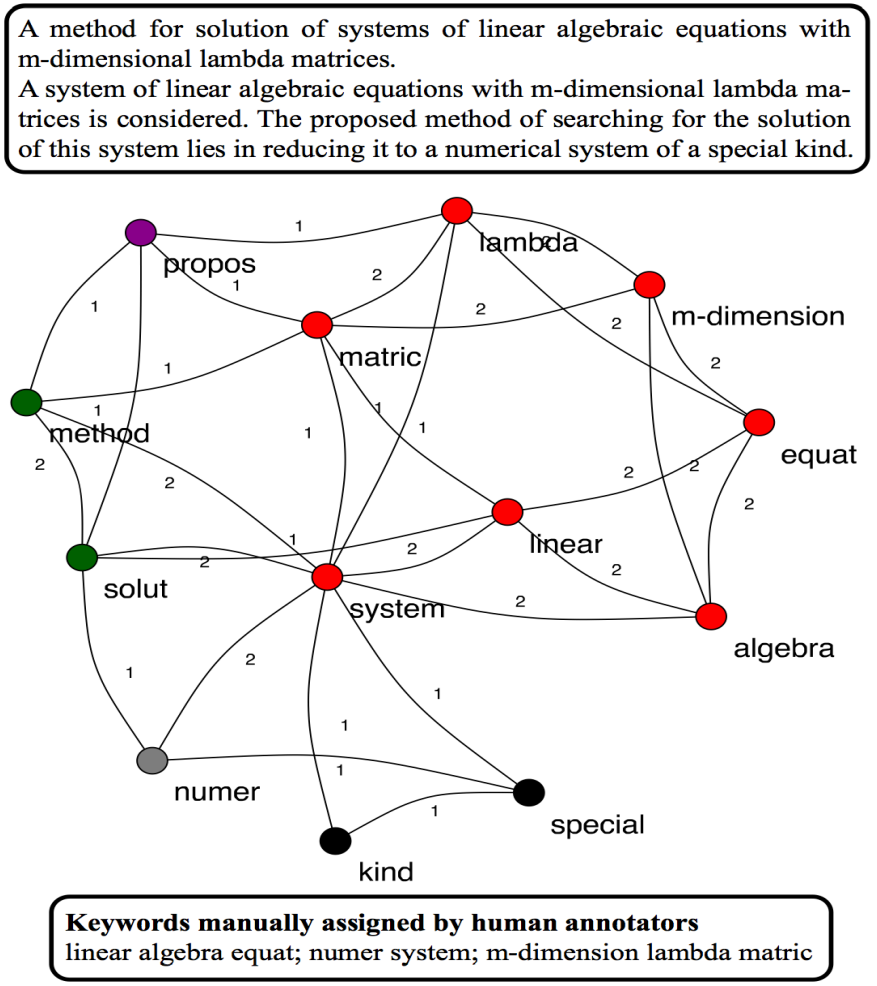


k-core decomposition of the graph

Idea: retain the k-core subgraph of the graph to extract the nodes based on their centrality and cohesiveness

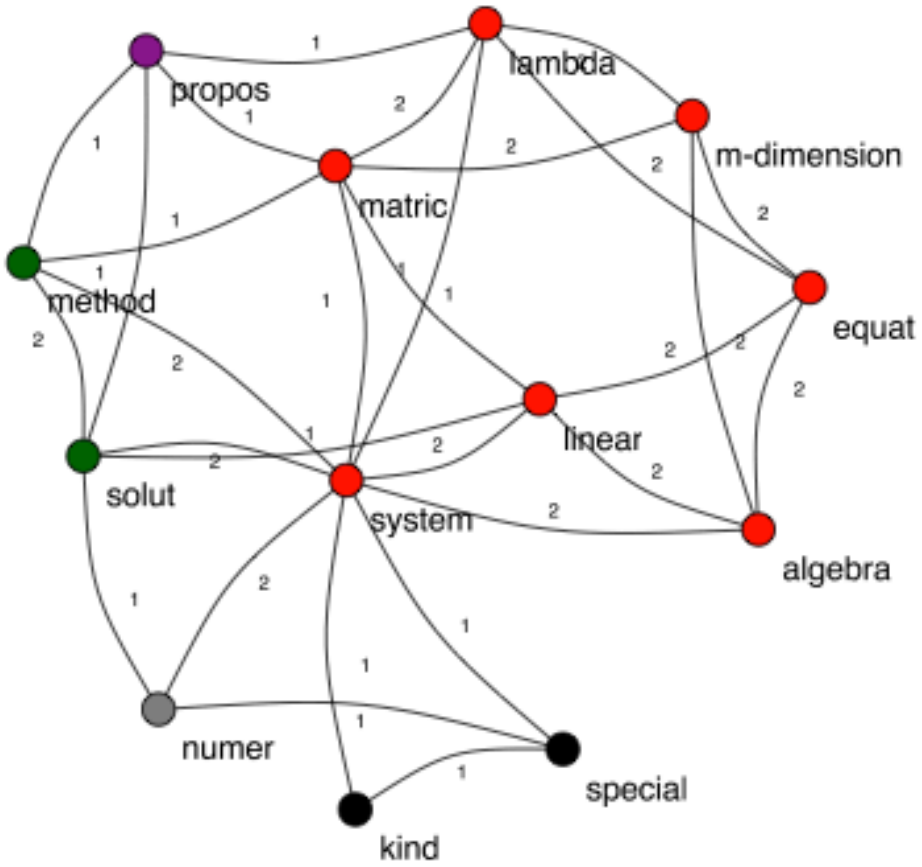
Graph-based Keyword Extraction (2/2)

- Single-document keyword extraction
  - Select the most cohesive sets of words in the graph as keywords
  - Use k-core decomposition to extract the main core of the graph
  - Weighted edges



[Rousseau and Vazirgiannis, ECIR '15]

PageRank vs. k-core



Keywords manually assigned by human annotators
linear algebra equat; numer system; m-dimension lambda matric

WK-core		PageRank	
system	6	**system**	1.93
matric	6	**matric**	1.27
lambda	6	solut	1.10
linear	6	**lambda**	1.08
equat	6	**linear**	1.08
algebra	6	**equat**	0.90
m-dim...	6	**algebra**	0.90
method	5	**m-dim...**	0.90
solut	5	propos	0.89
propos	4	method	0.88
numer	3	special	0.78
specia	2	**numer**	0.74
kind	2	kind	0.55

Keywords are not Unigrams

- 500 abstracts from the *Inspec* database used in our experiments,
- 4,913 keywords manually assigned by human annotators
- only 662 are unigrams (13%).
- Bigrams (2,587 – 52%) … 7-grams (5).

⇒ keywords are bigrams, if not higher order n-grams.

⇒ the interactions within keywords need to be captured in the first place – i.e. in the graph.

⇒ we can consider a k-core to form a "long-distance (k+1)-gram" [Bassiou and Kotropoulos, 2010]

How Many Keywords?

- Most techniques in keyword extraction assign a score to each feature and then take the top ones
- But how many?
  - Absolute number (top **X**) or relative number (top **X%**)?
- Besides, at fixed document length, humans may assign more keywords for a document than for another one

X is decided at document level (size of the k-core subgraph)

k-cores are adaptive

Datasets

- ***Hulth2003*** – 500 abstracts from the *Inspec* database [Hulth, 2003]

- ***Krapi2009*** – 2,304 ACM full papers in Computer Science (references and captions excluded) [Krapivin et al., 2009]

All approaches are unsupervised and single-document

Models and Baseline Methods

Graph-of-words:

- Undirected edges
- Forward edges
  - Natural flow of the text
  - An edge ***term1*** → ***term2*** meaning that ***term1*** precedes ***term2*** in a sliding window
- Backward edges

Keyword extractors:

- PageRank
- HITS (authority scores only)

} Top 33% or top 15% keywords

- k-core
- Weighted k-core

} Main core

Evaluation Metrics

- Each document has a set of golden keywords assigned by humans
  - **precision**, **recall** and **F1-score** per document
  - **macro-average** each metric at the collection level

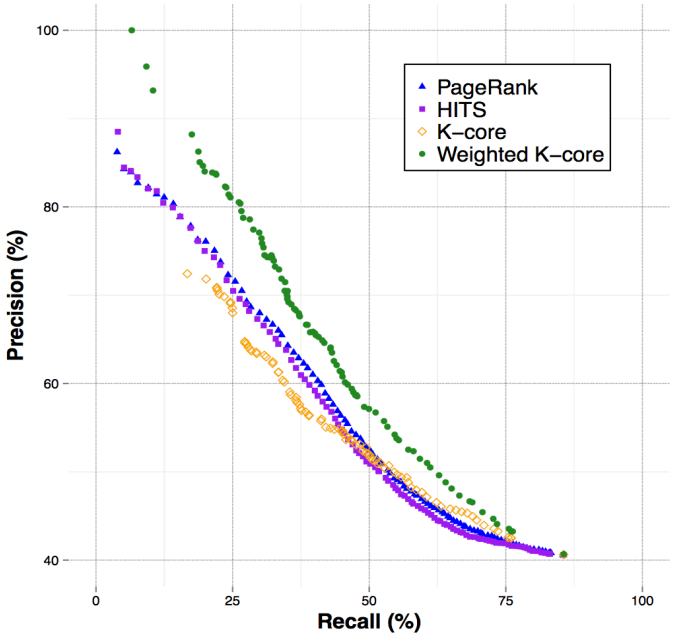
Performance Evaluation

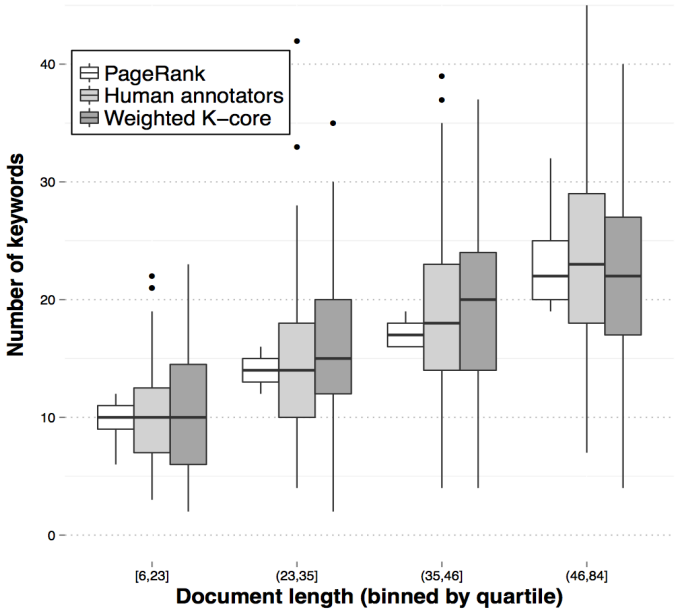
Precision

Recall

F1-score

Precision/recall





Graph	Dataset	Macro-averaged precision (%)				Macro-averaged recall (%)				Macro-averaged F1-score (%)			
		PageRank	HITS	K-core	WK-core	PageRank	HITS	K-core	WK-core	PageRank	HITS	K-core	WK-core
undirected edges	Hulth2003	58.94	57.86	46.52	**61.24***	42.19	41.80	**62.51***	50.32*	47.32	46.62	49.06*	**51.92***
	Krapi2009	50.23	49.47	40.46	**53.47***	48.78	47.85	**78.36***	50.21	49.59	47.96	46.61	**50.77***
forward edges	Hulth2003	55.80	54.75	42.45	**56.99***	41.98	40.43	**72.87***	46.93*	45.70	45.03	**51.65***	50.59*
	Krapi2009	47.78	47.03	39.82	**52.19***	44.91	44.19	**79.06***	45.67	45.72	44.95	46.03	**47.01***
backward edges	Hulth2003	59.27	56.41	40.89	**60.24***	42.67	40.66	**70.57***	49.91*	47.57	45.37	45.20	**50.03***
	Krapi2009	51.43	49.11	39.17	**52.14***	49.96	47.00	**77.60***	50.16	**50.51**	47.38	46.93	50.42

Example – ECIR'15 Paper

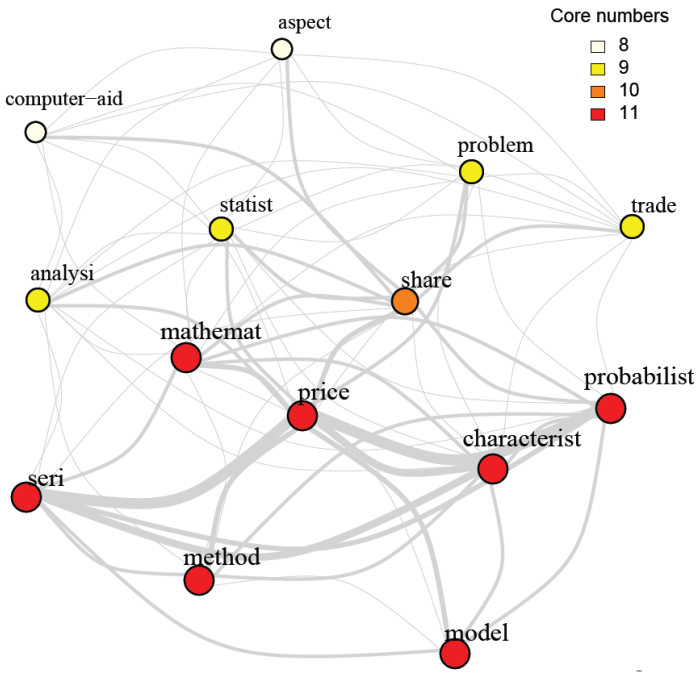
- Stemmed unigrams of the main core of the graph- of-words of the paper document: {*keyword, extract, graph, represent, text, weight, graph-of-word, k-core, degeneraci, edg, vertic, number, document*}

- Using PageRank, "work" appears in the top 5, "term" and "pagerank" in the top 10, and "case" and "order" in the top 15. Central words but not in cohesion with the rest and probably not relevant

A Different Point of View

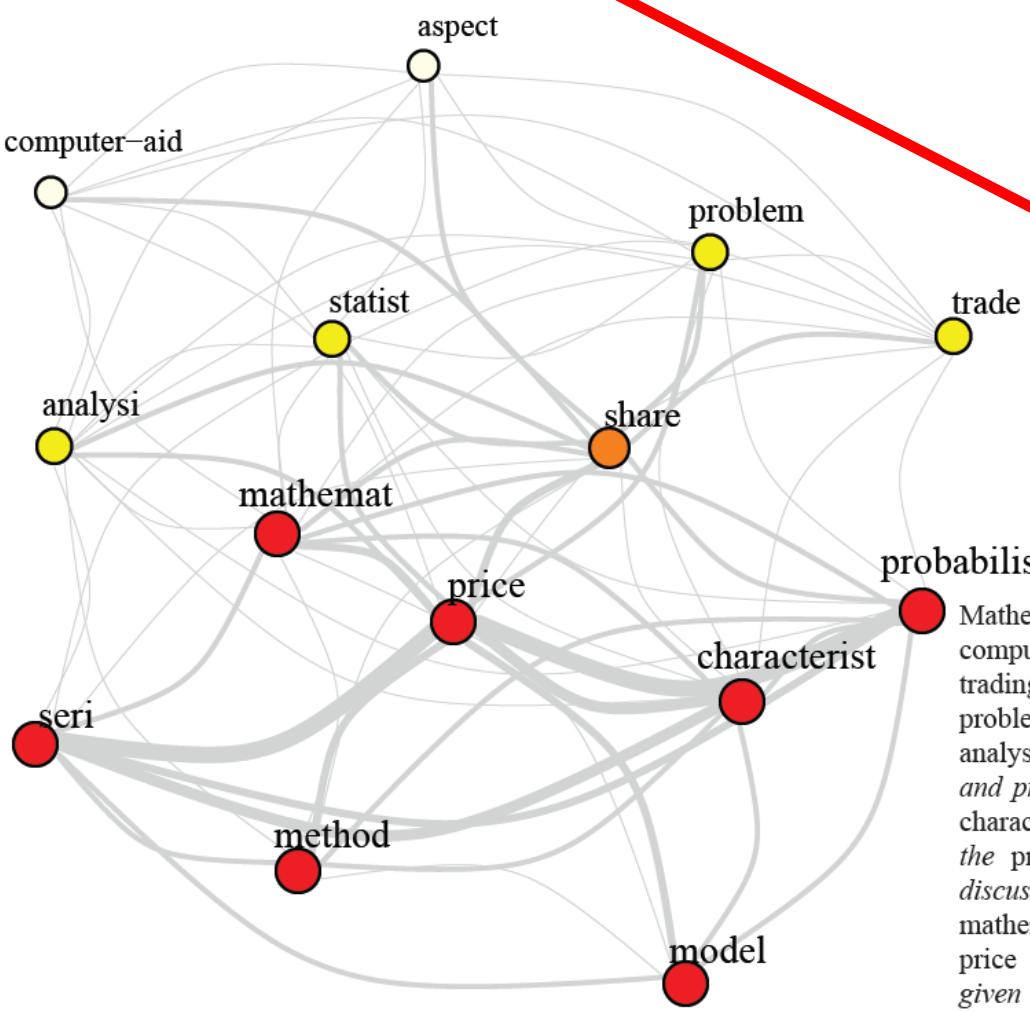
Graph degeneracy:

- In social networks, nodes part of the highest levels of the hierarchy are better spreaders than nodes high on PageRank
- Nodes with high truss numbers are even more influential than nodes with high core numbers
- **Spreading influence** may be a better "keywordness" metric than prestige (captured by PageRank)



Drawbacks of Graph Degeneracy (1/4)

Retaining the top level like in may be an appealing initial idea



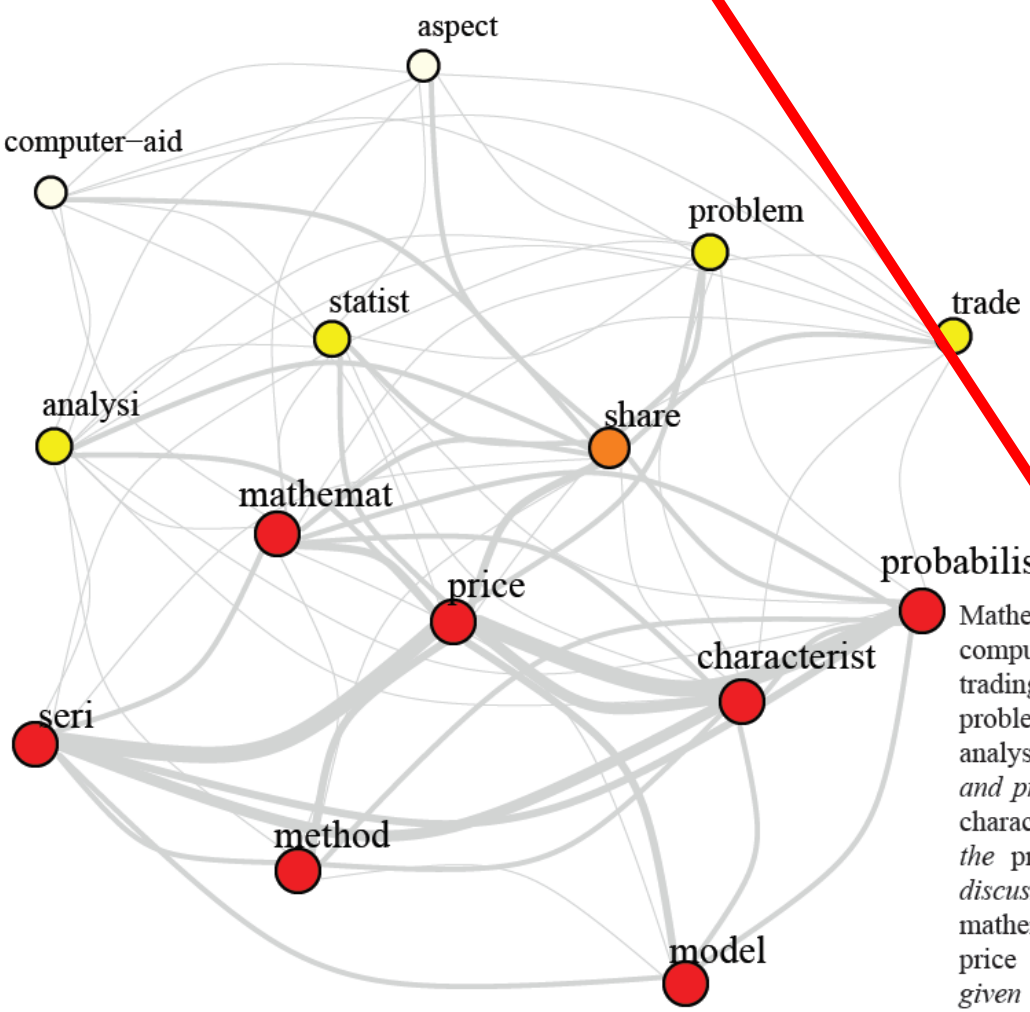
Mathematical aspects *of* computer-aided share trading. *We consider* problems *of* statistical analysis *of* share prices *and propose* probabilistic characteristics *to describe the* price series. *We discuss three* methods *of* mathematical modelling *of* price series *with given* probabilistic characteristics.

Core numbers			
	P	R	F1
MAIN	0.86	0.55	0.67
INF	**0.83**	**0.91**	**0.87**
DENS	0.88	0.64	0.74

mathemat		11
price		11
probabilist		11
characterist		11
seri		11
method		11
model	*MAIN*	11
share	*DENS*	10
trade		9
problem		9
statist		9
analysi	*INF*	9
aspect		8
computer-aid		8

Drawbacks of Graph Degeneracy (2/4)

But many keywords live below the top level -> good precision, poor recall



Edge weights
1
2
3
4
5

Core numbers
8
9
10
11

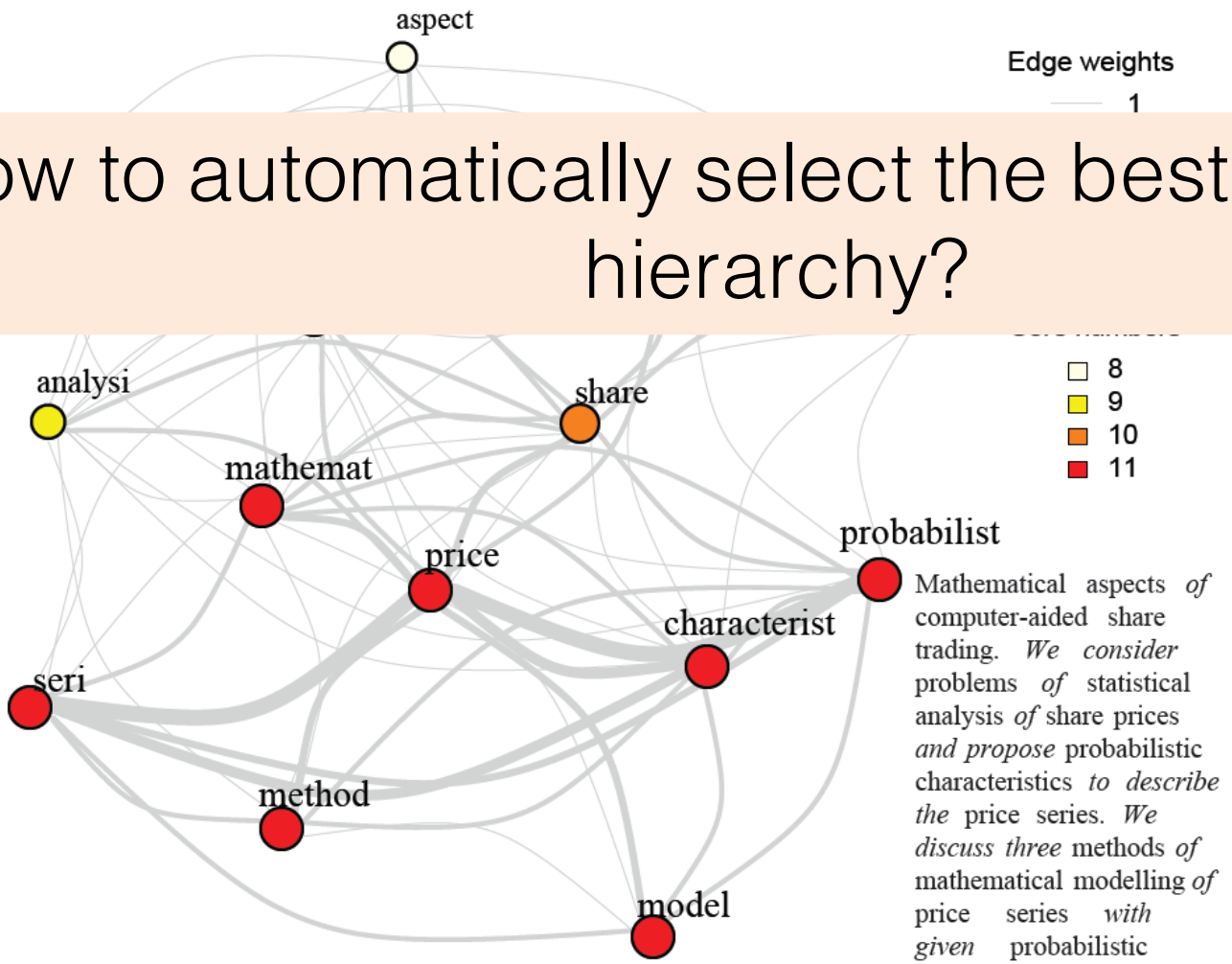
Mathematical aspects *of* computer-aided share trading. *We consider* problems *of* statistical analysis *of* share prices *and propose* probabilistic characteristics *to describe the* price series. *We discuss three* methods *of* mathematical modelling *of* price series *with given* probabilistic characteristics.

Core numbers			
	P	R	F1
MAIN	0.86	0.55	0.67
INF	**0.83**	**0.91**	**0.87**
DENS	0.88	0.64	0.74

mathemat		11
price		11
probabilist		11
characterist		11
seri		11
method		11
model	*MAIN*	11
share	*DENS*	10
trade		9
problem		9
statist		9
analysi	*INF*	9
aspect		8
computer-aid		8

Drawbacks of Graph Degeneracy (3/4)

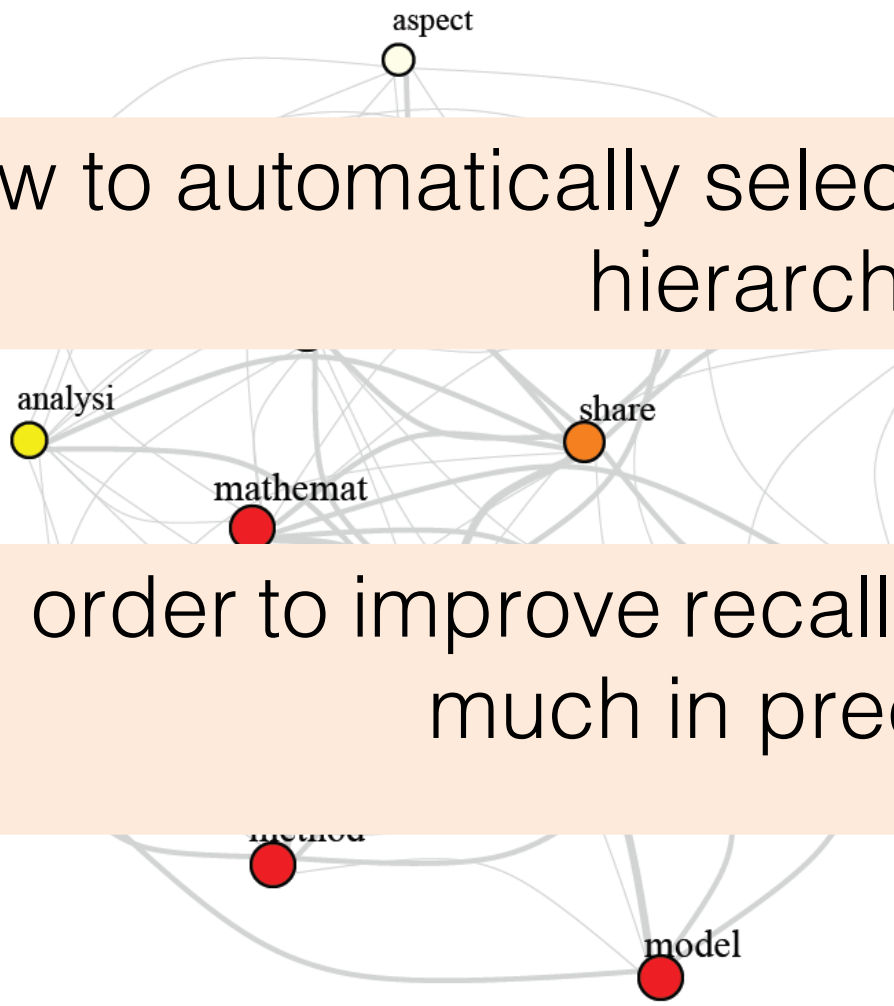
How to automatically select the best level in the hierarchy?

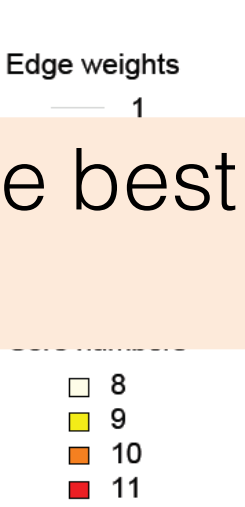


Mathematical aspects *of* computer-aided share trading. *We consider* problems *of* statistical analysis *of* share prices *and propose* probabilistic characteristics *to describe the* price series. *We discuss three* methods *of* mathematical modelling *of* price series *with given* probabilistic characteristics.

Core numbers			
	P	R	F1
probabilist			11
characterist			11
seri			11
method			11
model	*MAIN*		11
share	*DENS*		10
trade			9
problem			9
statist			9
analysi	*INF*		9
aspect			8
computer-aid			8

Drawbacks of Graph Degeneracy (4/4)





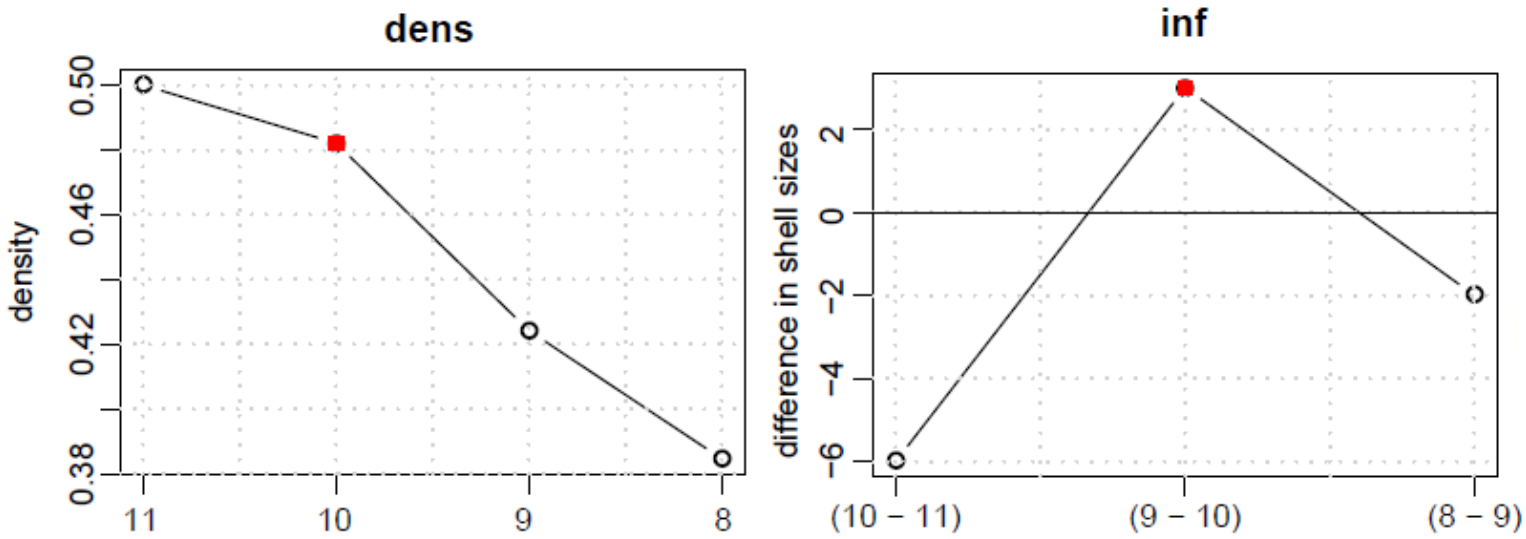
Core numbers			
	P	R	F1
probabilist			11
characterist			11
seri			11
statist			9
analysi		*INF*	9
aspect			8
computer-aid			8

How to automatically select the best level in the hierarchy?

In order to improve recall while not losing too much in precision?

the price series. *We discuss three* methods *of* mathematical modelling *of* price series *with given* probabilistic characteristics.

Graph Degeneracy for Keyword Extraction



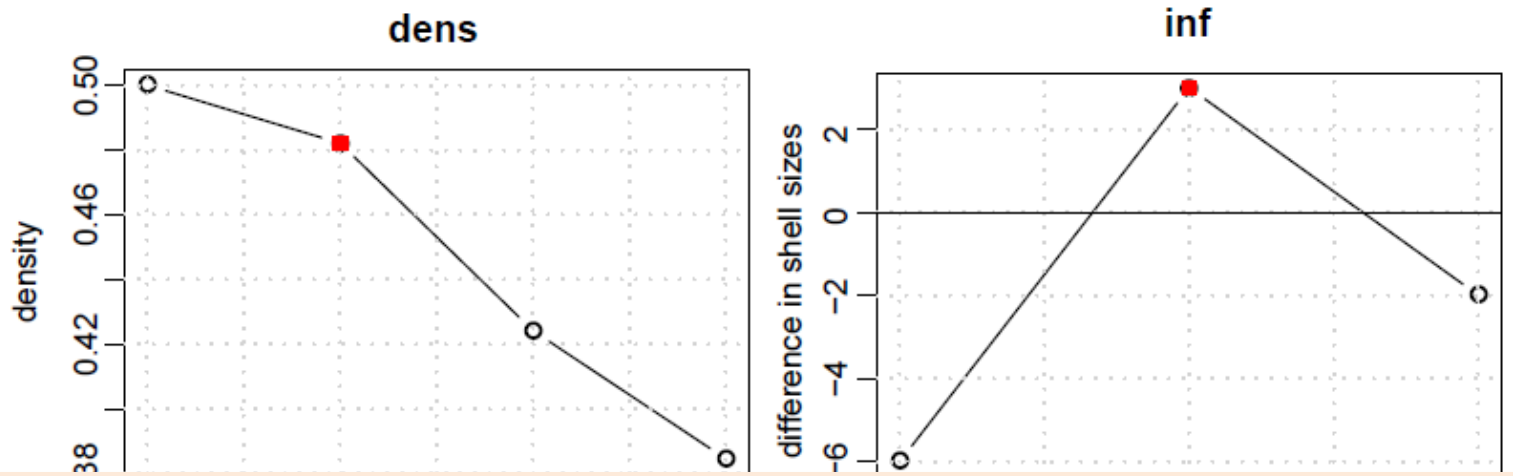
Heuristics:

- dens: go down the hierarchy until a drop in k-core (or truss) density is observed, i.e., as long as the desirable cohesiveness properties are kept
- inf: go down the hierarchy as long as the shells increase in size (starting at the main – 1 level)

Problem: both methods work at the subgraph level -> lack flexibility for large graphs (adding an entire group of nodes or not)

[Tixier et al., EMNLP '16]

Graph Degeneracy for Keyword Extraction



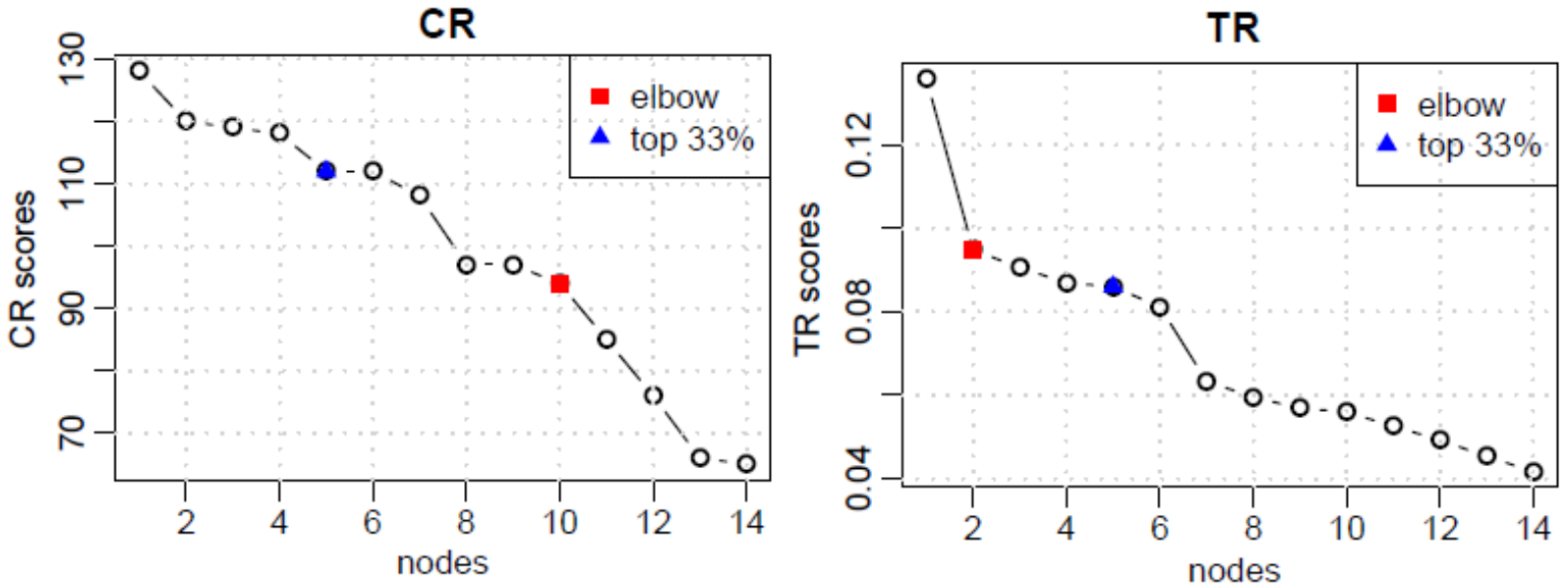
How to work at the node level while still retaining the valuable cohesiveness information captured by degeneracy?

- inf: go down the hierarchy as long as the shells increase in size (starting at the main – 1 level)

Problem: both methods work at the subgraph level -> lack flexibility for large graphs (adding an entire group of nodes or not)

[Tixier et al., EMNLP '16]

CoreRank



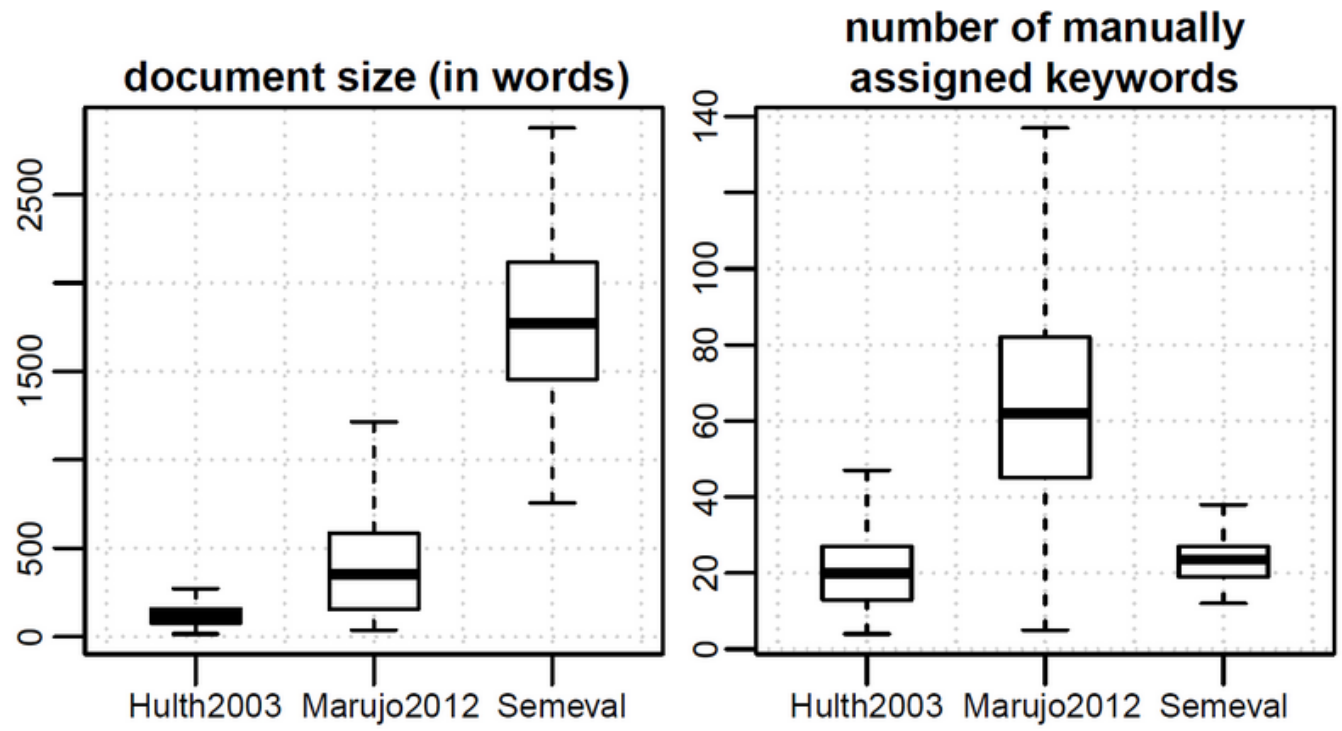
CoreRank (CR):

- Assign to each node the sum of the core (or truss) numbers of its neighbors
- Granularity is much finer and allows for much flexible selection
- Comparable to applying PageRank to the graph-of-words (aka TextRank) but taking into account cohesiveness concerns rather than individual prestige only

Heuristics: nodes can be selected based on the elbow or top p% method

[Tixier et al., EMNLP '16]

CoreRank – Experimental Evaluation (1/2)



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

CoreRank – Experimental Evaluation (2/2)

	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†

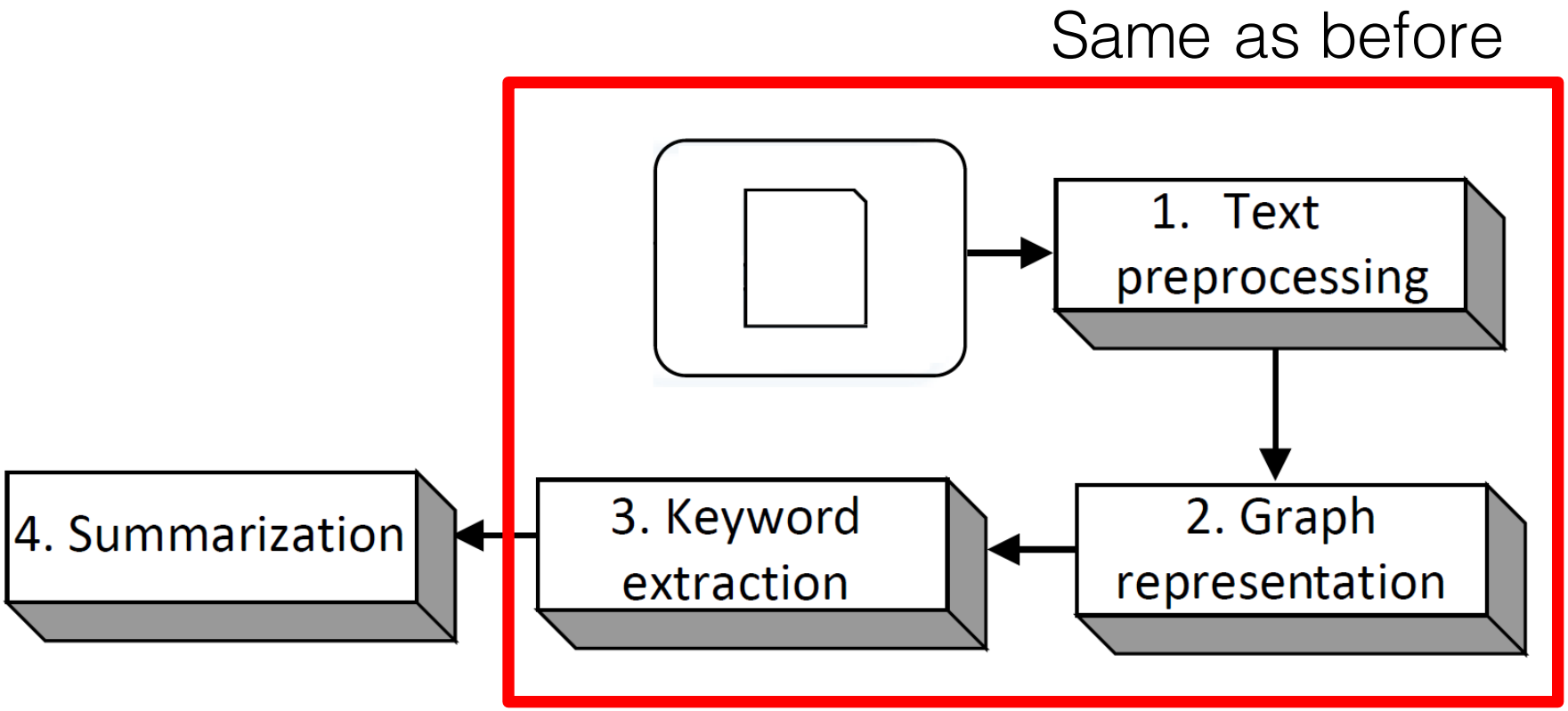
	precision	recall	F1-score
dens	8.44	79.45	15.06
inf	17.70	65.53	26.68
CRP	**49.67**	**32.88**	**38.98***
CRE	25.82	58.80	34.86
main†	25.73	49.61	32.83
TRP†	47.93	31.74	37.64
TRE†	33.87	46.08	37.55

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

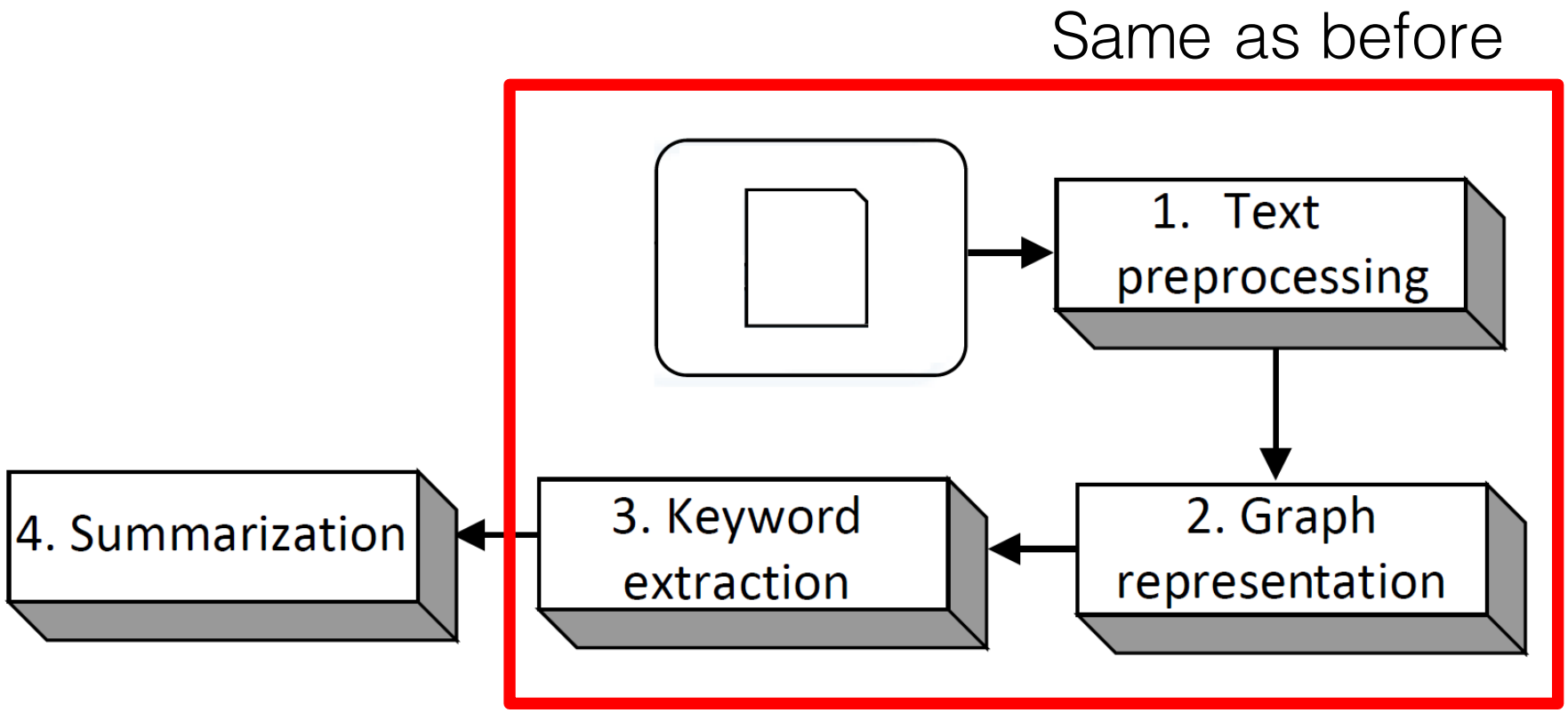
- For small documents (i.e., small graphs), the subgraph-level heuristics significantly outperform main core retention (main) and TextRank (TRP, TRE)
- Recall is drastically improved, precision is maintained (especially with inf)
- For long documents (Semeval), the node-level heuristics are better
- CoreRank with top p% retention (CRP) reaches best performance

Extractive summarization

Extension to Extractive Document Summarization



Extension to Extractive Document Summarization



How to use keywords (and their scores) to select the best sentences in a document?

Extractive Document Summarization (1/4)

- Generating a summary in an extractive way is akin to selecting the best sentences in the document under a budget constraint (max number of words allowed)
- Combinatorial optimization task:

$$\arg\max_{S \subseteq V} F(S) \mid \sum_{v \in S} c_v \leq B$$

  - **S** is a given summary (a subset of the set of sentences **V**)
  - **F** is the objective function to maximize (measuring summary quality)
  - $\mathbf{C_v}$ is the cost of sentence **v** (number of words it contains)
  - **B** is the budget (in words)

[Lin and Bilmes, NAACL '10]

Extractive Document Summarization (2/4)

$$\arg\max_{S \subseteq V} F(S) \mid \sum_{v \in S} c_v \leq B$$

- Solving this task is NP-complete
- It has been shown that if **F** is non-decreasing and submodular, a greedy
  algorithm can approach the best solution with factor **(e – 1)/e**
- At each step, the algorithm selects the sentence v that maximizes:

objective function gain →

$$\frac{F(G \cup v) - F(G)}{c_v^r}$$

scaled cost →

- **r** is a tuning parameter

Extractive Document Summarization (3/4)

$$\arg\max_{S \subseteq V} F(S) \mid \sum_{v \in S} c_v \leq B$$

- The choice of **F**, the summary quality objective function, is what matters
- A good summary should cover all the important topics in the document, while not repeating itself
  - Maximize **coverage**
  - Penalize **redundancy** (reward diversity to ensure monotonicity)

$$F(S) = L(S) + \lambda R(S)$$

$$L(S) = \sum_{i \in S} n_i w_i \qquad R(S) = N_{keywords \in S} / N_{keywords}$$

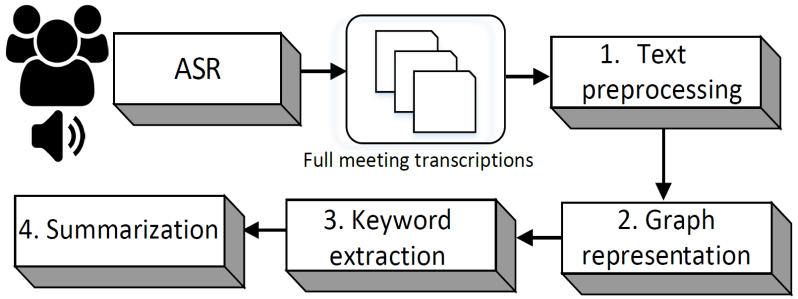
weighted sum of the keywords contained in the summary

proportion of unique keywords contained

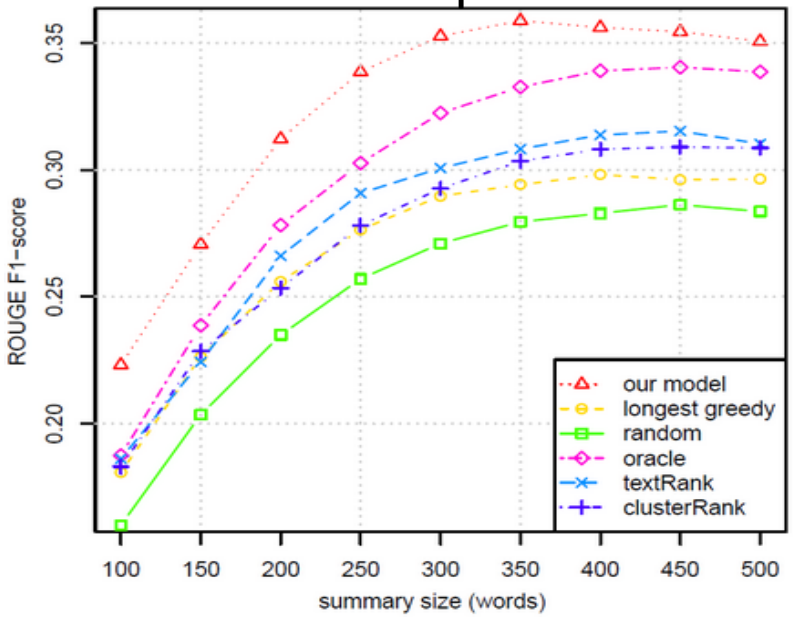
[Lin and Bilmes, NAACL '10], [Meladianos et al., EACL '17]

Extractive Document Summarization (4/4)

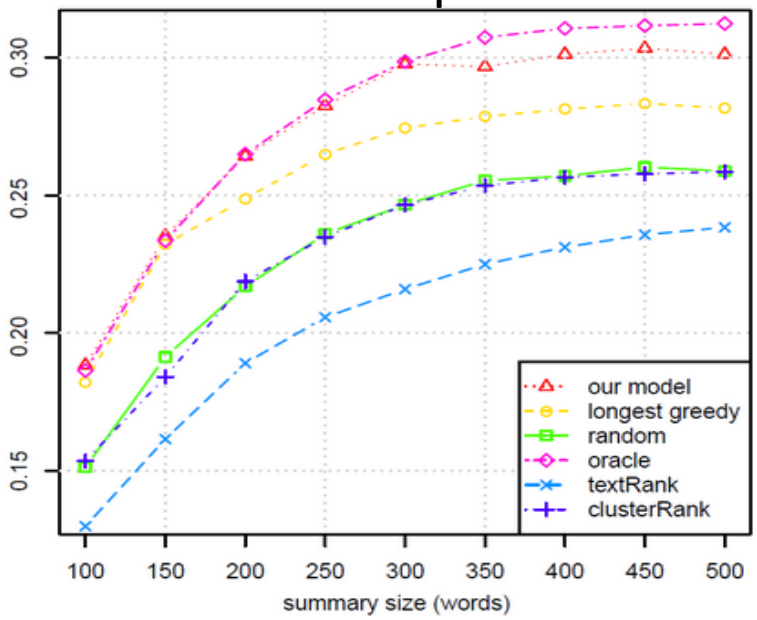
Tested for multiparty virtual meetings summarization:



AMI corpus



ICSI corpus



Event detection in text streams
1: Threshold based approach

Introduction - Twitter

- Very popular microblogging service
  - 330 million active users
- Real time messages known as tweets
  - Instantaneous nature
  - Major communication medium
- Why Twitter
  - Large variety of usages: communication, news, politics, celebrities
  - Huge volume of data
  - Transmission in real time

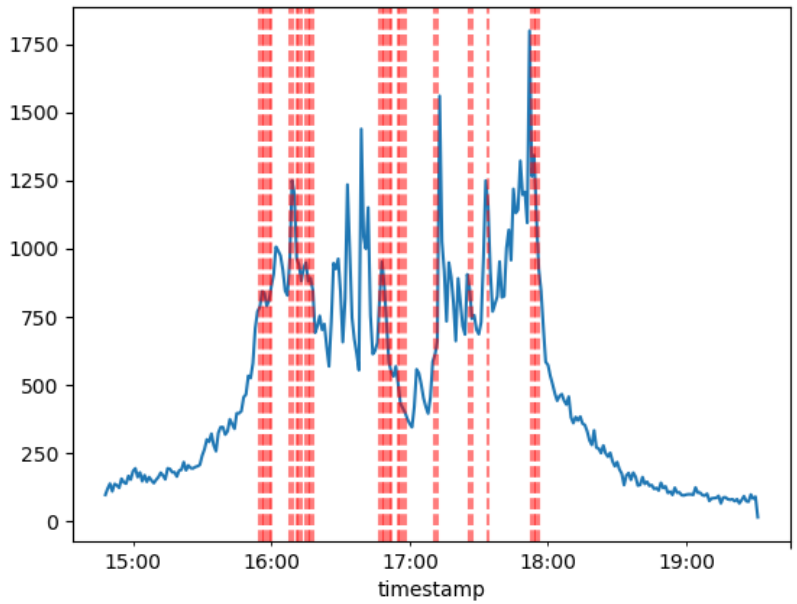
Introduction - Event Detection

- Users report latest news or comment about real-world events
  - Events consisting of a sequence of important moments
  - Events not covered systematically by traditional media
  - Interest in tracking the evolution of an event

- Challenges:
  - Huge volume of tweets - Noisy content
  - Heterogeneous users
  - Multiple hashtags per event
  - Generating short summary

Sub-event Detection in Twitter Streams

1. Large volume of documents in social medi
2. Events are not covered by traditional medi
3. News appear fast in Twitter
4. Is Tweet rate suited for sub-event Detection?



- Tweet Rate of a football Match
- Red lines: True Events
- Very noisy during the event!

Tweet Rate histogram of a football match

[Meladianos et al., AAAI-ICWSM '15]

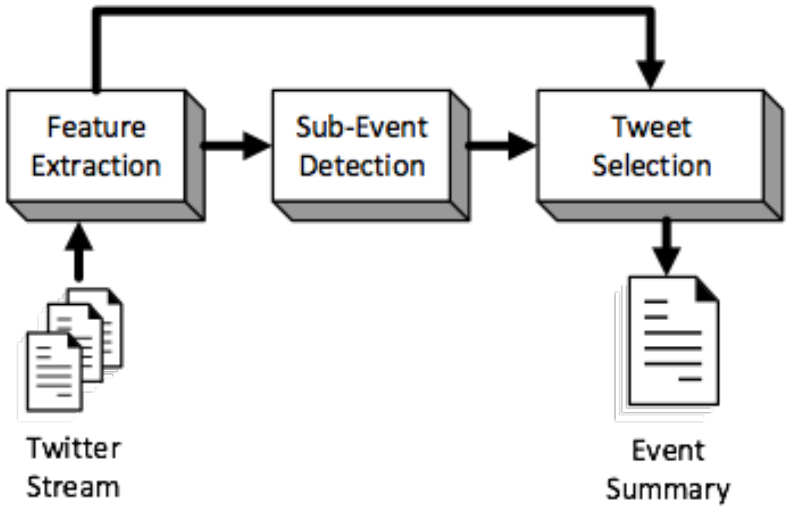
Sub-event Detection in Twitter Streams

Real time event summarization

1. Feature extraction: extracts the terms that best describe the current state of the event

1. Sub-event detection: decides whether a sub-event has occurred

1. Tweet selection: ranks all the tweets and selects the first one



System Architecture

- Steps are repeated every 60 seconds
- The summary of the whole event is constructed by aggregating the individual sub-event descriptions

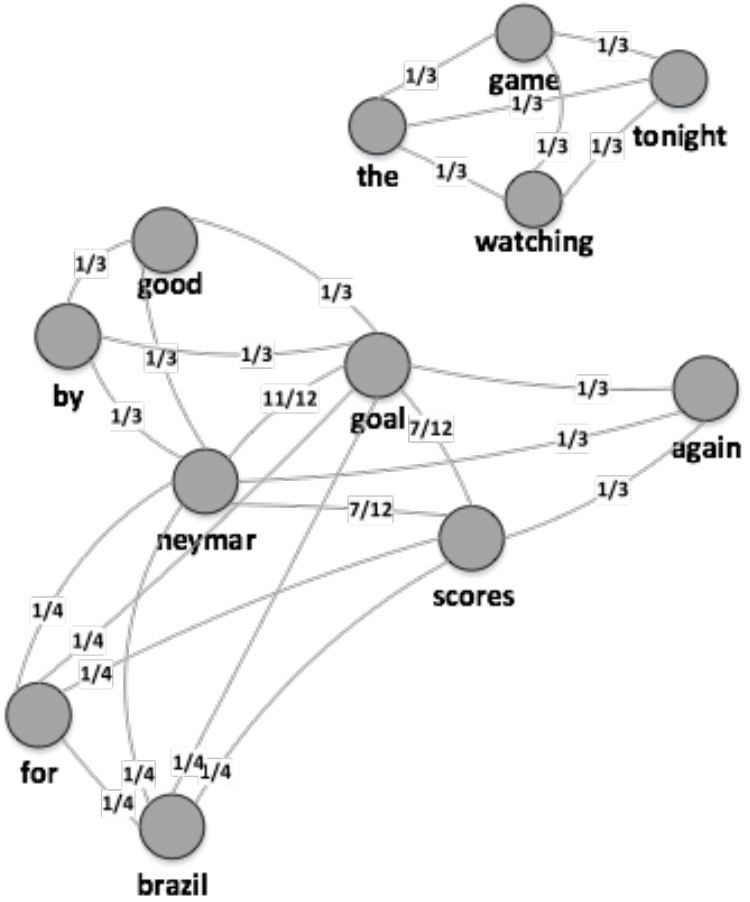
Graph-based Representation of Tweets

- Represents all the input tweets
- Nodes: unique terms
- Edges: #co-occurrences within a tweet

Example graph

1. Good goal by Neymar
2. Goal! Neymar scores for brazil
3. Goal!! Neymar scores again
4. Watching the game tonight

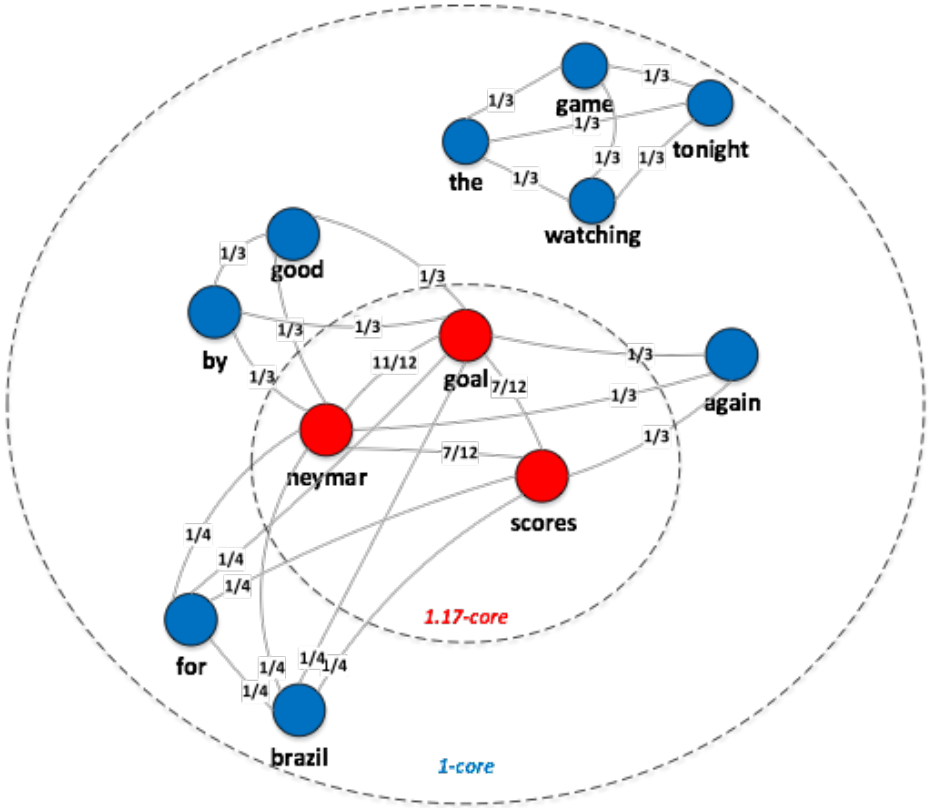
Dataset: tweets from the 2014 FIFA World Cup in Brazil



The graph that was built from 4 tweets

k-core Decomposition for Feature Extraction

- Each term is given a score corresponding to its core number
- Extract the k-core subgraph
- Detect sub-events by considering how the sum of the core numbers extracted from the graph at time **t** has changed from a previous time point **t-1**



k-core decomposition of the Graph-of-Words

Sub-event Detection

$$\sum_{i=1}^{d} c_i^t > \theta \times \frac{1}{p} \sum_{j=t-p}^{t-1} \sum_{i=1}^{d} c_i^j$$

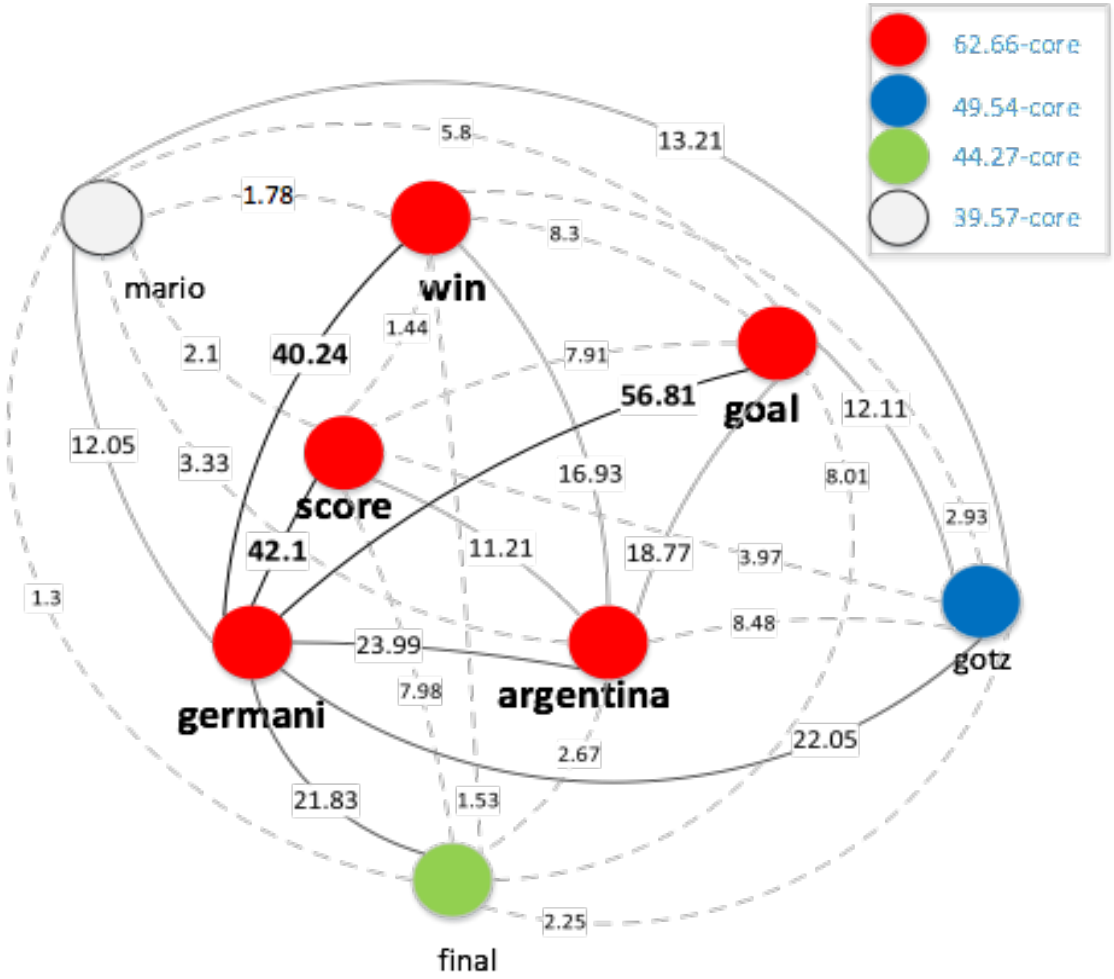
c_i^t	Core number of term at time slot
d	Number of terms selected
θ	Decision Threshold
p	Number of previous time slots

Sub-event Detection steps:

(every 60 seconds)

1. Extract the top terms with highest weights
2. Sum the term weights
3. If it exceeds the threshold a sub-event is detected

Germany's Goal - 2014 World Cup

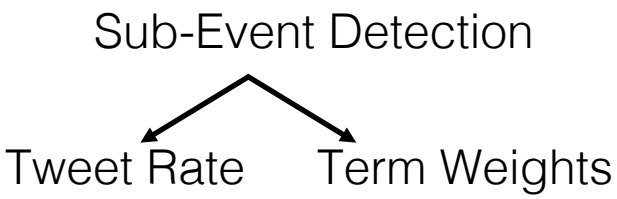


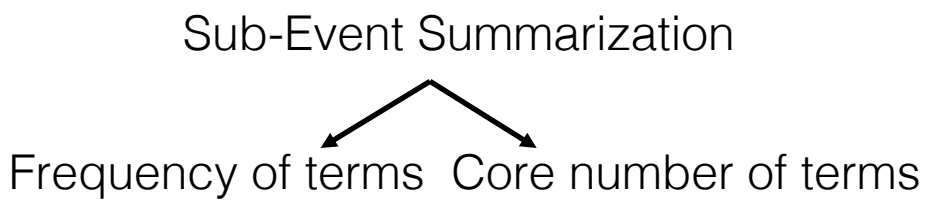
Snapshot of the four highest cores of the graph generated after Germany's goal in the 2014 FIFA World Cup final

Tweet Selection as Sub-event Summarization

- Activated only if a sub-event has been detected
- Tweets are scored based on the sum of their term weights
- Selects the most informative tweet of the sub-event
  - The tweet with the highest score is chosen

Exprimental Setup





Baselines-Approaches

(Detection -Term Weight)

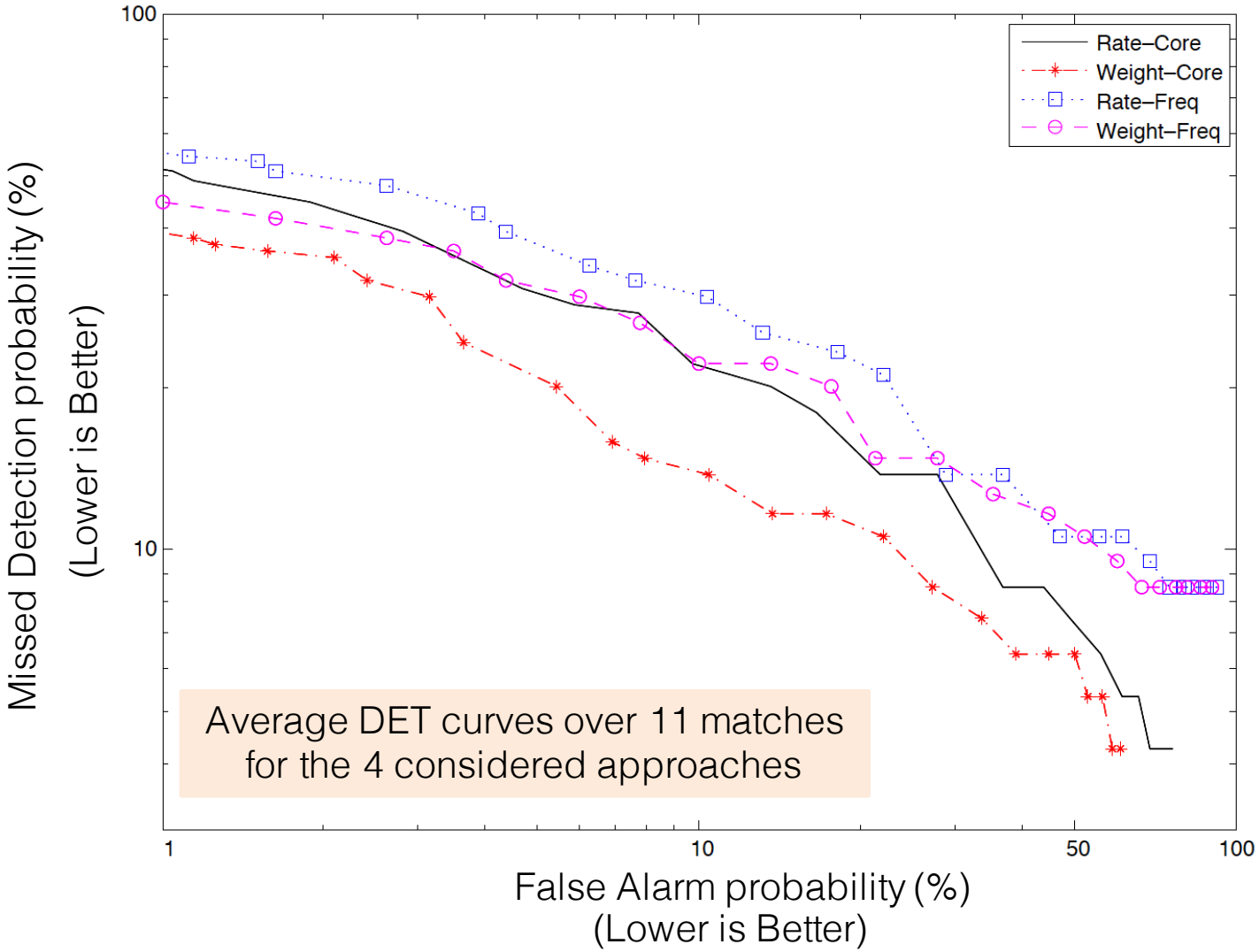
- ➢ Rate–Freq: the common baseline
- ➢ Rate–Core
- ➢ Weight–Core: Our approach
- ➢ Weight–Freq

Dataset

Match	#sub-events	#tweets
Germany - Argentina	8	1,907,999
Argentina - Belgium	7	1,355,472
France - Germany	6	1,321,781
Honduras-Switzerland	7	168,519
Greece - Ivory Coast	10	251,420
Croatia - Mexico	11	600,776
Cameroon - Brazil	11	532,756
Netherlands - Chile	7	301,067
Australia - Spain	9	252,086
Germany - Ghana	8	718,709
Australia - Netherlands	11	126,971
All Matches	95	7,537,556

FIFA 2014 World Cup Dataset

Evaluation (1/2)



Evaluation (2/2)

Method	Micro F1-score	Macro F1-score
Weight-Core	0.68	0.72
Rate-Core	0.61	0.63
Weight-Freq	0.61	0.64
Rate-Freq	0.54	0.60

Average micro and macro F1-score over 11 matches for the 4 considered approaches

Event type	#actual Events	#detected Events
Goal	32	30
Penalty	2	2
Red Card	1	0
Yellow Card	27	14
Match Start	11	8
Match End	11	11
Half Time	11	10

Number of sub-events detected

Tweet Summarization Performance

Time	Summary	ESPN FC
8’	Goal!!!!Argentina!! After eight minutes Argentina lead Belgium by 1-0 scored by Higuain	Goal! Argentina 1, Belgium 0. Gonzalo Higuain (Argentina) right footed shot from the centre of the box to the bottom left corner.
45’+2’	HT: Argentina 1-0 Belgium. Fantastic goal by Higuain gives Argentina the slight lead over the red devils.	First Half ends, Argentina 1, Belgium 0.
52’	52m - Belgium's Eden Hazard with the first yellow card of the game	Eden Hazard (Belgium) is shown the yellow card for a bad foul.
75’	Argentina 1 - 0 Belgium | Biglia booked a yellow card. Meanwhile, Chadli on for Eden Hazard.	Lucas Biglia (Argentina) is shown the yellow card for a bad foul.
90+5'	Well at least that goal makes them advance to the semi finals. Argentina gets the ticket to advance and Belgium goes home.	Match ends, Argentina 1, Belgium 0.

Summary of the Argentina vs. Belgium match generated automatically using Weight–Core and manually by ESPN

Summary

- Sub-event detection approach based on the k-core decomposition on graph-of-words
- The algorithm exploits the fact that the vocabulary of tweets gets more specific when a sub-event occurs
- The detection mechanism is able to accurately detect important moments as they occur
- The tweets selected by our system give an overview of the event

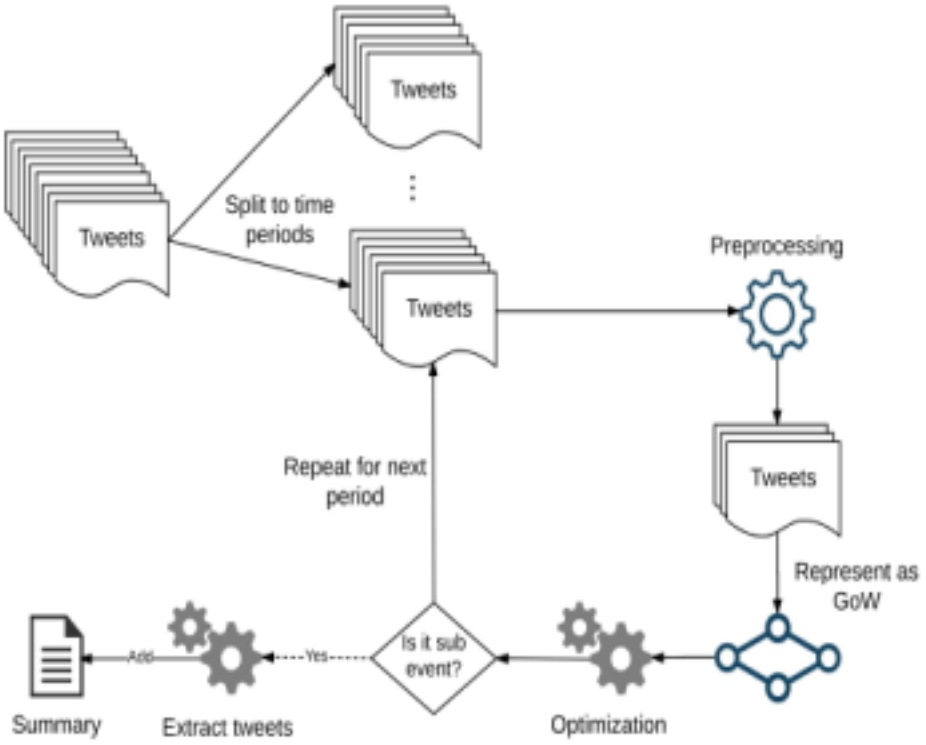
[Meladianos et al., AAAI-ICWSM '15]

Event detection in text streams

2: An Optimization Approach for Sub-event Detection and Summarization in Twitter

Introduction - System Architecture

- Tweets decomposition into time intervals
- Graph of Words representation
- Convex optimization formulation
- Summarization



Methodology - Data Preprocessing

Given a set of raw tweets:

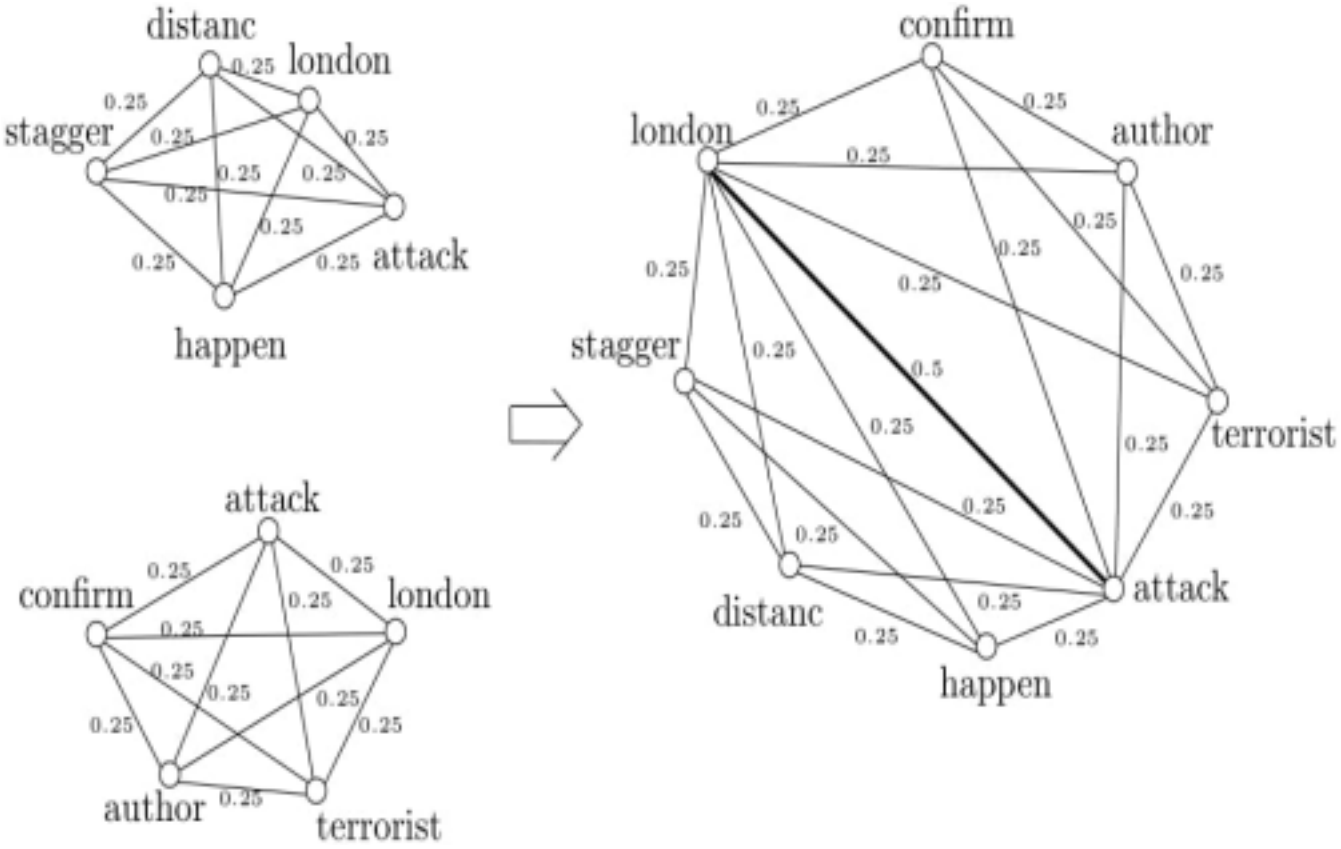
- Remove not relevant tweets
  - Retweets
  - Duplicates
  - Mentions
- Text preprocessing:
  - Tokenization
  - Stopword removal
  - Punctuation, special character and URL removal
  - Stemming using Porter's algorithm

Methodology - Graph of Words

Given a set of pre-processed tweets:

- Each tweet represented as a weighted graph
  - Vertices correspond to unique terms
  - Edges are drawn between all pairs of vertices (clique)
  - Weight set equal to $1/(n-1)$ n : number of unique terms in the tweet
  - The degree of all vertices is equal to 1
- Creating a graph G_i corresponding to the time period i
  - Add tweet-graphs sequentially
  - Increase weight of edges occurring in multiple tweets

Methodology - Graph of Words



Methodology - Sub-Event Detection

Sub-Event detection:

- Change rate among tweets posted currently compared to previous time periods
- Pairs of words with high co-occurence frequency.

- Detection handled as *optimization problem*
  - For time period *i*, graph G_i
  - Adjacency Matrix $\mathbf{A}_i \in \mathbb{R}^{n \times n}$
  - Represents co-occurrence among words in tweets

$$\mathbf{A}_i = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & & | \end{bmatrix}$$

- *b* is the *n^2* rows matrix representing as well the co-occurrence of words in the tweets in the periods *1-p*

$$\mathbf{b} = vec(\mathbf{A}_i) = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}$$

Methodology - Sub-Event Detection

- Construct matrix W with weights from previous *p* periods
- each w_i represents the co-occurrence weights of all terms in the graph time i (summarizing previous tweets)

$$\mathbf{W} = \begin{bmatrix} | & & | \\ \mathbf{w}_{i-p} & \dots & \mathbf{w}_{i-1} \\ | & & | \end{bmatrix} \quad \mathbf{W} \in \mathbb{R}^{n^2 \times p}$$

- Optimization problem formulated as:
- x represents the vector that matches best the current to the previous p time
- Event: large values of objective function
  - when $OF > \theta$ event

$$\min_{\mathbf{x}} \frac{1}{2} ||\mathbf{W}\mathbf{x} - \mathbf{b}||_2^2$$
$$\text{s.t.} \quad \mathbf{1}^\top \mathbf{x} = 1$$
$$x_i \geq 0, \quad \forall i = 1, \dots, p$$

Methodology - Sub-Event Detection

Methodology - Summarization (1/2)

Extractive summarization algorithm → Select a subset of tweets

These tweets will serve as the summary of the sub-event

Goal: Selected tweets containing most significant terms of the detected sub-event

- Define a monotone sub modular function f that rewards both the **coverage** and the **diversity** of the set
- proposed function f can identify multiple sub-events in a single time period

Methodology - Summarization (2/2)

A sub-event is best described by tweets whose graph-of-words contains multiple "important" edges

G_i : graph-of-words representation of tweets posted during time period i

$\mathcal{S} \subseteq D_i$: a set of tweets extracted from time period i

$f(\mathcal{S})$: the sum of weights of the edges of G_i that are "covered" by tweets belonging to $\mathcal{S}$

f is monotone non-decreasing and submodular → use greedy approximation algorithm

- guaranteed to be within $(1-1/e) \approx 0.63$ of the optimal solution

Evaluation - Experiments

Baseline method

- Tweeting rate (Burst)
  - Number of tweets in period
  - Ignores previous periods

Evaluation settings:

- Threshold: Training process
- Period duration: 1 min
- Number of previous periods: 10 periods

Evaluation - Dataset

- 18 FIFA World Cup 2014 and 2 FIFA World Cup 2010 matches
  - 3 used for tuning and the rest for testing
- Only some sub-event types considered
  - goals, penalties, cards, match start, match end and half time
  - injuries, substitutions, free kicks etc. ignored
- Groundtruth
  - Manual annotation of events

Match	Actual Events	Collected Tweets	Preprocessed Tweets
ARG - BEL	7	313,803	108,250
ARG - GER	9	824,241	262,112
AUS - NED	12	96,834	25,997
AUS - ESP	9	86,843	13,608
BEL - KOR	7	99,192	32,053
CMR - BRA	11	148,298	35,085
FRA - GER	6	525,725	160,727
FRA - NGA	6	367,899	128,718
GER - ALG	8	712,525	276,227
GER - BRA	12	973,985	295,875
GER - GHA	8	285,804	77,449
GER - USA	7	256,445	86,040
GRE - CIV	10	113,402	51,101
HON - SUI	8	41,539	10,082
MEX - CRO	11	155,549	36,981
NED - CHI	8	95,108	25,819
NED - MEX	10	628,698	217,472
POR - GHA	10	272,389	91,110
GER - SRB	14	45,024	29,062
USA - SVN	12	85,675	53,292
Total	185	6,128,978	2,017,060

Parameter Tuning

- Relationship between the optimal value of θ and the number of tweets
- Randomly selected 3 matches as training examples
  - 17 matches were placed into the test set
- Exhaustive grid search, $\theta \in [1, 100]$
- Minimize the least-squares-error

$$\hat{\theta} = w_1 x^2 + w_2 x + w_3$$

Evaluation - Results

Method \ Metric	Macro-average			Micro-average		
	precision	recall	f1-score	precision	recall	f1-score
OptSumm	0.76	0.75	0.75	0.73	0.74	0.73
Burst	0.78	0.54	0.64	0.72	0.54	0.62

- Evaluate the proposed system (OptSumm) on the task of sub-event detection
- Standard measures in information retrieval
  1) Precision
  2) Recall
  3) F1-score
- OptSumm detected sub-events that could not be detected by Burst,

Evaluation - Sub-events Detected

- Key sub-event types, their actual numbers and detected numbers over the 17 matches of the test set
- Successfully detected all goals, own goals, penalties and red cards
- Successfully detected all match ends and half times, and almost all match starts
- failed to detect the yellow cards consistently

Event type	# actual events	# detected events
Goal	42	42
Own Goal	2	2
Penalty	3	3
Red Card	3	3
Yellow Card	51	15
Match Start	17	14
Match End	17	17
Half Time	17	17

Evaluation - Summary

- System produces informative and reasonable textual descriptions of the important moments.

Our Summary	FIFA
Underdog Nigeria vs. European giants France. Going to be a great match!	The match kicks off.
Nigeria awarded a free kick in a good position after Matuidi collides with Odemwingie. Nigeria looking decent on the break so far. #fra #nga	Matuidi (France) concedes a free-kick following a challenge on Odemwingie (Nigeria).
France #fra 0-0 Nigeria #nig - Nigeria with Eminike score, but ruled out for offside, good decision 18mins	Emenike (Nigeria) is adjudged to be in an offside position.
Pogbaaaaaa!!! excellent skill! made that entire move and ended it with a superb volley but keeper made a good save	Pogba (France) sees his effort hit the target.
Half time: France 0-0 Nigeria. Goalless in brasília. tight game.	The referee brings the first half to an end.
54: Blaise Matuidi gets the first yellow card of the game after a nasty challenge.	Matuidi (France) is booked by the referee.
Nigeria the best team by far. That usually means France will scrape a lucky win.	-
Omg! #Benzema so close to scoring, just 2? inches short. Still 0-0 (Nigeria-France) in a suddenly very exciting match!	Benzema (France) sees his effort hit the target.
If the French don't score in this game, it would be a miracle for Nigeria. France has been inches away from about 3 goals at this point.	-
Goal France! Who else, but the future, Paul Pogba, heading into an open net. Finally les blues score. 1-0, 80th min.	Pogba (France) scores!!
Goool! France scores in the 91st minute! Partial score, France 2-0 Nigeria. #worldcup goal count - 147.	Yobo (Nigeria) scores an own goal!!
Full-time: #fra 2-0 #nga. France book their spot in the quarter-final while Nigeria crash out of the 2014 Fifa world cup	The final whistle sounds.

Evaluation - Other events

- Westminster bridge (London)
  - Terrorist attack on 22 March 2017
  - 5 people died, at least 50 were injured

- Collected 153, 485 tweets
  - Time period equal to 15 minutes
  - Taking into consideration only the previous period

Evaluation - Other events

Timestamp	Description
22 Mar 14:56	#BREAKING London Terror Attack - Parliament lockdown, man with knife shot by security.Many people ran down by car on Westminster Bridge. ◊ REPORTS AT LEAST 12 INJURED IN A CAR RAMMING/STABBING ATTACK OUTSIDE PARLIAMENT IN LONDON, UK. ASSAILANT SHOT BY POLICE.
22 Mar 15:26	#DEVELOPING: London police are treating the attack at Westminster near British Parliament "as a terrorist incident until we know otherwise.". ◊ #BREAKING Met confirm terror attack.Officer stabbed, shots heard at Parliament London.Many injured on Westminster Bridge hit by vehicle.
22 Mar 17:56	4 people dead, including police officer, and at least 20 injured in Parliament attack #Parliament London#terrorism. ◊ Thoughts with the victims of today's terrorist attack, their families & all those caught up at #Westminster. London will #KeepCalmAndCarryOn.
22 Mar 22:26	Metropolitan Police say 5 people have died and around 40 have been injured in the Westminster terror attack in London today #londonattack. ◊ 4 people including a terrorist and a police officer have been killed and 20 others injured in an terror attack in London.
23 Mar 10:26	Seven arrested as police raid addresses in Birmingham and London linked to Westminster terror attack #York. ◊ Mashable "Donald Trump Jr. tweeted about the London attack and got instantly dragged" #News.
23 Mar 11:56	BREAKING: Islamic State claims responsibility for terror attack outside Parliament in London #Westminster #londonattack. ◊ News agency linked to ISIS is out with a claim of responsibility from the terrorist organization for yesterday's deadly attack in #London.
23 Mar 15:26	Breaking:London police identify man behind UK terror attack as 52-year Khalid Masood, a Briton with a criminal history. ◊ London attack: Westminster terrorist named as Khalid Masood by Scotland Yard after Isil claims responsibility.

Conclusion

- Presented a system capable of:
  - Generating real-time summaries
  - Detecting sub-events
  - Works in different kinds of events

Summary

- Graphs have been widely used as modeling tools in
  - NLP
  - Text Mining
  - Information Retrieval

- Goal of the tutorial
  - Presentation of recent methods that rely on graph-based text representations to deal with various tasks in NLP and IR
  - Focus on the graph-of-words model
  - Borrow ideas from the graph mining and network analysis field

References (1/4)

- Boudin, F., and Morin, E. 2013. Keyphrase Extraction for N-best reranking in multi-sentence compression. In North American Chapter of the Association for Computational Linguistics (NAACL).
- Mehdad, Y., Carenini, G., Tompa, F. W., and Ng, R. T. 2013. Abstractive meeting summarization with entailment and fusion. In Proc. of the 14th European Workshop on Natural Language Generation (pp. 136-146).
- Manu Aery and Sharma Chakravarthy. 2005. Infosift: Adapting graph mining techniques for text classification. In FLAIRS, pages 277–282.
- Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto.1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Roi Blanco and Christina Lioma. 2012. Graph-based term weighting for information retrieval. Inf. Retr., 15(1):54–92.
- Florian Boudin. 2013. A comparison of centrality measures for graph-based keyphrase extraction (ijcnlp '13). In Sixth International Joint Conference on Natural Language Processing, pages 834–838.
- Adrien Bougouin, Florian Boudin, and B´eatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In Sixth International Joint Conference on Natural Language Processing (IJCNLP '13), pages 543–551.
- Adrien Bougouin, Florian Boudin, and B´eatrice Daille. 2016. Keyphrase annotation with graph co-ranking. In 26th International Conference on Computational Linguistics (COLING '16).
- Gunes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, pages 457–479.
- Filippova, K. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 322-330). Association for Computational Linguistics.

References (2/4)

- Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. 2009. Extracting key terms from noisy and multitheme documents. In Proceedings of the 18th International Conference on World Wide Web (WWW '09) , pages 661–670. ACM.
- Samer Hassan, Rada Mihalcea, and Carmen Banea. 2007. Random-walk term weighting for improved text classification. In ICSC, pages 242–249.
- Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito. 2010. Text classification using graph mining-based feature extraction. Knowl.- Based Syst., 23(4):302–308.
- Marina Litvak and Mark Last. 2008. Graphbased keyword extraction for single-document summarization. In Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization (MMIES '08), pages 17–24. Association for Computational Linguistics.
- Marina Litvak, Mark Last, Hen Aizenman, Inbal Gobits, and Abraham Kandel. 2011. Degext — a language-independent graph-based keyphrase extractor. In Elena Mugellini, Piotr S. Szczepaniak, Maria Chiara Pettenati, and Maria Sokhn, editors, Proceedings of the 7th Atlantic Web Intelligence Conference (AWIC '08), pages 121–130.
- Fragkiskos D. Malliaros and Konstantinos Skianis. 2015. Graph-based term weighting for text categorization. In ASONAM, pages 1473–1479. ACM.
- Alex Markov, Mark Last, and Abraham Kandel. 2007. Fast categorization of web documents represented by graphs. In Advances in Web Mining and Web Usage Analysis, volume 4811, pages 56–71.
- Polykarpos Meladianos, Giannis Nikolentzos, Franc¸ois Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2015. Degeneracy-based real-time subevent detection in twitter stream. In Ninth International AAAI Conference on Web and Social Media (ICWSM '15).

References (3/4)

- Rada Mihalcea and Paul Tarau. 2004. TextRank: bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP '04). Association for Computational Linguistics.
- Francois Rousseau and Michalis Vazirgiannis. 2013. Graph-of-word and tw-idf: new approach to ad hoc ir. In Proceedings of the 22nd ACM international conference on Conference on Information & Knowledge Management (CIKM '13), pages 59–68. ACM.
- Francois Rousseau and Michalis Vazirgiannis. 2015. Main core retention on graph-of-words for singledocument keyword extraction. In European Conference on Information Retrieval (ECIR '15), pages 382–393. Springer.
- Francois Rousseau, Emmanouil Kiagias, and Michalis Vazirgiannis. 2015. Text categorization as a graph classification problem. In ACL (1), pages 1702–1712. The Association for Computer Linguistics.
- Francois Rousseau. 2015. Graph-of-words: mining and retrieving text with networks of features. Ph.D. thesis, Ecole Polytechnique.
- Stephen B. Seidman. 1983. Network Structure and Minimum Degree. Social Networks, 5:269–287.
- Konstantinos Skianis, Francois Rousseau, and Michalis Vazirgiannis. 2016. Regularizing text categorization with clusters of words. In EMNLP, pages 1827–1837. The Association for Computational Linguistics.
- Shashank Srivastava, Dirk Hovy, and Eduard H. Hovy. 2013. A walk-based semantically enriched tree kernel over distributed word representations. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13), pages 1411–1416.
- Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In Proceedings of the 24th International Conference on World Wide Web (WWW '15), pages 1067-1077

References (4/4)

- Antoine J.-P. Tixier, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. 2016a. A graph degeneracybased approach to keyword extraction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16), pages 1860–1870. The Association for Computational Linguistics.
- Antoine J.-P. Tixier, Konstantinos Skianis, and Michalis Vazirgiannis. 2016b. Gowvis: a web application for graph-of-words-based text visualization and summarization. In ACL. The Association for Computational Linguistics.
- S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. 2010. Graph kernels. J. Mach. Learn. Res., 11:1201–1242.
- Jia Wang and James Cheng. 2012. Truss decomposition in massive networks. Proc. VLDB Endow., 5(9):812–823.
- Wei Wang, DiepBich Do, and Xuemin Lin. 2005. Term graph model for text classification. In Advanced Data Mining and Applications, volume 3584, pages 19–30.
- Rui Wang, Wei Liu, and Chris McDonald. 2015. Corpus-independent generic keyphrase extraction using word embedding vectors. In Workshop on Deep Learning for Web Search and Data Mining (DL-WSDM '15).
- Xifeng Yan and Jiawei Han. 2002. gSpan: Graphbased substructure pattern mining. In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02).

References – Event detection

- **Tweeting rate**: Volume of status updates
  - Nichols, J., Mahmud, J., Drews, C.: Summarizing Sporting Events Using Twitter. In: IUI. pp. 189–198 (2012)
  - Zhao, S., Zhong, L., Wickramasuriya, J., Vasudevan, V.: Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games. arXiv:1106.4300 (2011)
  - Zubiaga, A., Spina, D., Amigó, E., Gonzalo, J.: Towards Real-time Summarization of Scheduled Events from Twitter Streams. In: HT. pp. 319–320 (2012)
- **Retweet rate**: Volume of retweets updates
  - Chierichetti, F., Kleinberg, J., Kumar, R., Mahdian, M., Pandey, S.: Event Detection via Communication Pattern Analysis. In: ICWSM. pp. 51–60 (2014)

References – Event detection

- **Modified Hidden Markov Model**: tweet rate and the word distribution
  - Chakrabarti, D., Punera, K.: Event Summarization Using Tweets. In: ICWSM. Pp. 66–73 (2011)
- **Mixture model**: detect sub-events for each participant
  - Shen, C., Liu, F., Weng, F., Li, T.: A Participant-based Approach for Event Summarization Using Twitter Streams. In: NAACL-HLT. pp. 1152–1162 (2013)
- **Hierarchical Dirichlet processes**: probabilistic topic model
  - Srijith, P., Hepple, M., Bontcheva, K., Preotiuc-Pietro, D.: Sub-story detection in twitter with hierarchical dirichlet processes. Information Processing & Management (2016**)**
- **Graph degeneracy**: sequences of tweets as graphs
  - Meladianos, P., Nikolentzos, G., Rousseau, F., Stavrakas, Y., Vazirgiannis, M.: Degeneracy-based Real-Time Sub-Event Detection in Twitter Stream. In: ICWSM.pp. 248–257 (2015)

Thank You! - Questions?

- **Michalis Vazirgiannis**
  École Polytechnique, France
  mvazirg@lix.polytechnique.fr
  http://www.lix.polytechnique.fr/~mvazirg