Who are we now?

An attempt to understand the ICDAR community!

Andreas Dengel

What makes a conference memorable?

(1) Find a conference place with a 'wow factor'

(2) Provide a excellent service to the audience

(3) Incorporate workshops that require involvement, i.e. encouraging senior and junior people to contribute in an open atmosphere

(4) Makes sure that the reception and the banquet are exceptional and top the expectations of the participants

(5) Invite keynote speakers who are engaged, convincing, entertaining, and talk about topics which are beyond traditional approaches

From a historical point of view, we considered a document image as a subject of study and interpretation

A document is a …

- a piece of paper, booklet, etc., providing information, especially of an official or legal nature — **paper**
- a piece of text or text and graphics stored on a computer as a file for manipulation by document processing software — **digital**
- evidence or a proof — **legal**

- a written or drawn representation of thoughts — **paper**
- a textual file along with its structure and design (fonts, colors, and additional images) — **digital**
- a written proof used as evidence — **legal**

Document Analysis and **Recognition*** includes contributions dealing with computer recognition of characters, symbols, text, lines, graphics, images, handwriting, signatures, as well as automatic analyses of the overall physical and logical structures of documents, **with the ultimate objective of a high-level understanding of their semantic content.**

* Sources: Wikipedia; Dictionary; International Journal on Document Analysis and Recognition (IJDAR)

The analysis task may not be restricted to a single document only but to an entire set of documents capturing important information in their combination

A documentation is a set of documents provided on paper, or online, or on digital or analog media

- It is mainly required to transfer information through time
- It is becoming less common to see paper (hard-copy) documentation
- User manuals, quick-reference guides, or conference proceedings are typical examples

Professionals educated in this field are termed documentalists. This field changed its name to **Information Science** in 1968, but some uses of the term documentation still exists and there have been efforts to reintroduce the term documentation as a field of study

* Sources: Wikipedia

Document Analysis and Recognition has a long tradition culminating in the foundation of the ICDAR conference series documented within the set of proceedings

- The proceedings corpus provides is a valuable source for a whole bunch of quantitative and qualitative analyses
- We may apply OCR and layout analysis on the entire corpus for reediting, reformatting, rearranging, etc. for, e.g. generate a book on specific topic or a historical overview
- We may analyze the contents to classify the various papers, to get an idea about the employment of specific approaches, etc.

The analysis may be also expanded one step more to get insights into the social network behind the proceedings corpus

In general, a social network is a social structure made of **nodes** (which are generally individuals or organizations) **that are tied by** one or more specific types of **interdependency**, such as values, visions, ideas, financial exchange, friends, kinship, dislike, conflict, trade, web links, sexual relations, disease transmission, or airline routes. The resulting structures are often very complex

Corpus captures a large-scale but hidden "social (semantic) structure with

- directed weighted relations,
- network properties,
- performances of subsets, and
- location of actors,

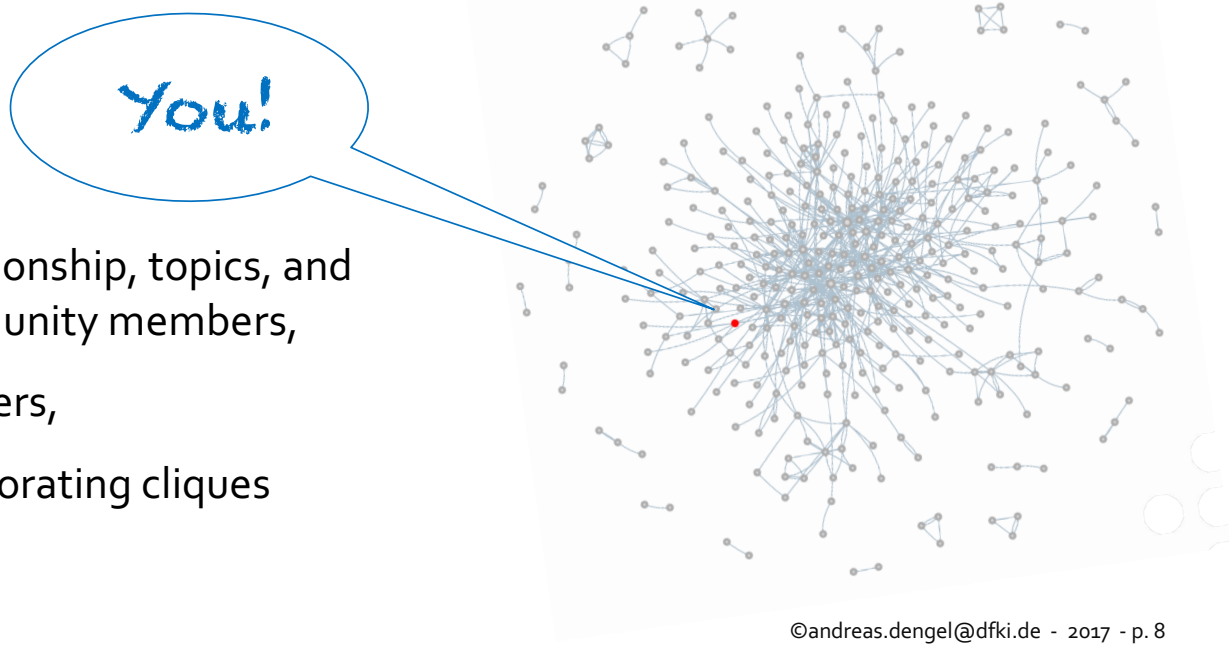
based on counting the incoming and outgoing links and applying mathematical models

Generally three basic types of networks can be created from a set of scientific publications

- **Co-authorship networks** is used as an indicator for the collaboration of authors (and their affiliated institutions)
  - Nodes represent the author
  - Node size conveys the number of publications by the author
  - Edge label denotes the number of times two authors have co-authored a paper
- **Publication citation networks** show relationships among scientific articles based on their citations
  - Direct citation network
  - A bibliographic coupling network where an edge is drawn between two publications if both cite the same previous publication(s)
  - Co-citation network where an edge between two publications exist if they both cite each others publication
- **Semantic networks** are formed on the basis of occurrence of a keyword in a set of publications (nodes represent words and edges represent the co-occurence of these words in one article)

This way, we approached this problem by a holistic document analysis approach to identify relations and to measure impact factors

Community Corpus → Syntactic Analysis → Semantic Analysis → Pragmatic Analysis → Academic Community Explorer



The goal is to reveal

- … historical contributors, their relationship, topics, and technologies addressed by community members,
- … scientific trends and opinion leaders,
- … citation behavior as well as collaborating cliques

Syntactic Document Analysis addresses the physical composition of a document

PDF

Preprocessing

1. Identify format (single/multiple columns)
2. Ignore tables, images
3. Transform text (image) into Unicode
4. Normalize the format of the references

Meta Data Extraction

1. Extract Meta Data (title, header, keywords, abstract and references)
2. Extract sentences capturing citations/references using Named Entity Recognition

Structured Data

- Citation Data (from 1993 – 2015)
- Collaboration Data (joined publication)
- Keyword Information (original pdf or IEEE Bibtex)

XML


```
<xml>
  <title>A Tool for Tuning Binarization Techniques</title>
  <Author>
    Vavilis Sokratis Information and Communication Systems Engineering
    University of the Aegean Samos , Greece sokratisvav@gmail.com Ergina
    Kavallieratou Information and Communication Systems Engineering
    University of the Aegean Samos , Greece kavallieratou@aegean.gr
  </Author>
  <Abstract>
    AbstractIn this paper a user friendly tool appropriate to get user
    feedback for the application of binarization algorithms is presented.
    The human feedback is very useful in order to apply next the algorithm
    to similar images. The tool supports Image Selection and Display ,
    Selection of Binarization Algorithm and Parameter Configuration ,
    Feedback gathering and Creation of log file for further processing.
  </Abstract>
  <Keywords>
    Keywords document image processing; binarization algorithms;user
    feedback
  </Keywords>
  Perfect reference extraction
  <References>
    [1] Jie Zou and George Nagy , Visible models for interactive pattern
    recognition , Pattern Recognition Letters 28 (2007) 2335–2342 [2]
    George Nagy and Sriharsha Veeramachaneni , Adaptive and Interactive
    Approaches to Document Analysis , Springer , Machine Learning in
    Document Analysis and Recognition , Volume 90/2008 [3] A Kesidis , E
    Galiotou , B Gatos , A Lampropoulos , Ioannis Pratikakis , Ioanna
    Manolessou , Angela Ralli , Accessing the content of Greek historical
    documents Proceedings of The Third Workshop on AND [4] H. Ma and D.
    Doermann , Adaptive OCR with Limited User Feedback , 8th Int'l Conf.
    Document Analysis and Recognition (ICDAR) , 2005 , pp 814 818 .Marte
    A. Ramirez Ortegon , Raul Rojas , [5] Fanbo Deng , Zheng Wu , Zheng Lu
    , and Michael S. Brown , BinarizationShop: a user assisted software
    suite for converting old documents to black and white" , In
    Proceedings of the 10th annual [6] Pavlos Stathis , Ergina
    Kavallieratou and Nikos Papamarkos , An Evaluation Survey of
    Binarization Algorithms on Historical Documents" , IEEE proceedings of
    19th International Conference on [7] E. Kavallieratou , E. Stamatatos
    , Improving the quality of degraded document images" , IEEE
    proceedings of DIAL , pp. 340 349 , Second International Conference on
    Document Image [8] Roberto Paredes , Ergina Kavallieratou , Rafael
    Dueire Lins , "ICFHR 2010 Contest: Quantitative Evaluation of
    Binarization Algorithms , " 12th International Conference on Frontiers
    in Handwriting Recognition , pp. 733 736 , 2010.
  </References>
</xml>
```

Semantic Document Analysis deals with the interpretation of the document structure and content

XML

Network Generation

1. Build a co-authorship network for each document
2. Build a author citation network for each document
3. Aggregate the individual data and weight both graphs, i.e. nodes wrt (1) and edges wrt (2)

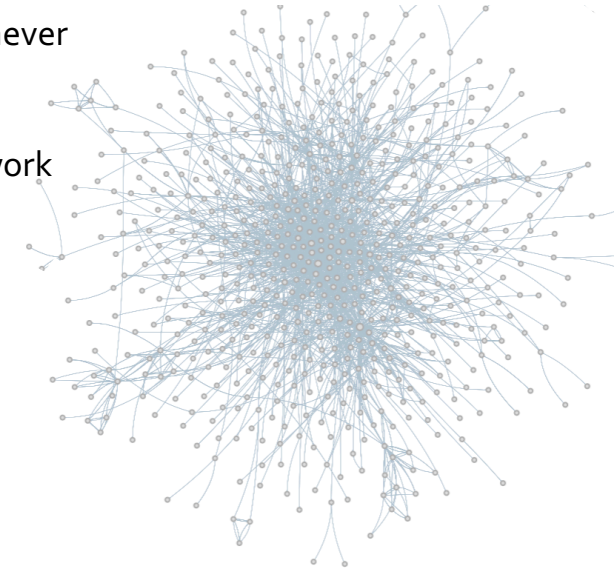
Data Cleaning and Polarity Measure

1. Delete "non-community" author names, i.e. authors who never published at ICDAR conferences
2. Name Resolution
3. Normalize the format of the author names and adapt network
4. Citation Sentiment Analysis

Structured Data

- B-PER, I-PER, E-PER tags for describing unique author names
- Aggregated collaboration and publication data

Enriched Networked Community Structure



Pragmatic Document Analysis tries to get insights into what is meant by the various propositions, i.e. its intent and effect

Enriched Networked Community Structure

Data Clustering

1. Clustering of co-authorship networks into groups of authors who frequently publish together

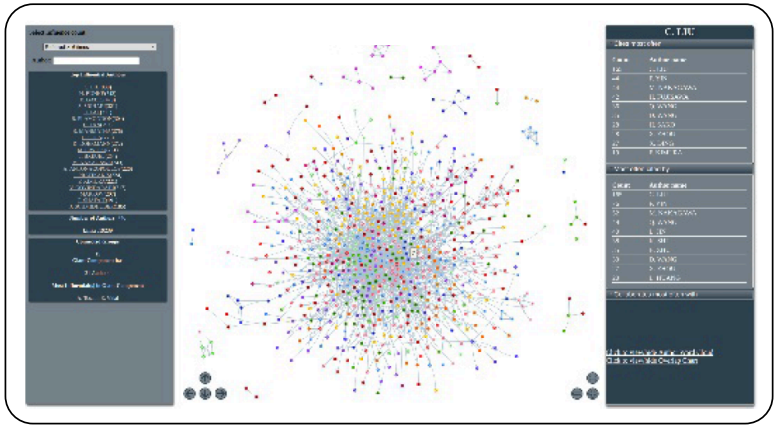
Performance Indicator Detection

1. Statistical summarization
2. Co-authorship evolvement analysis
3. Region-based co-authorship analysis
4. Measuring author centrality

Visualize Graph Data

- Intuitive representations of a large amount of data realizing multiscale navigations
- Interactive community data regarding authors, their collaboration and joint publishing
- Various statistics regarding the different roles of individuals, stratification, and cliques

Academic Community Explorer



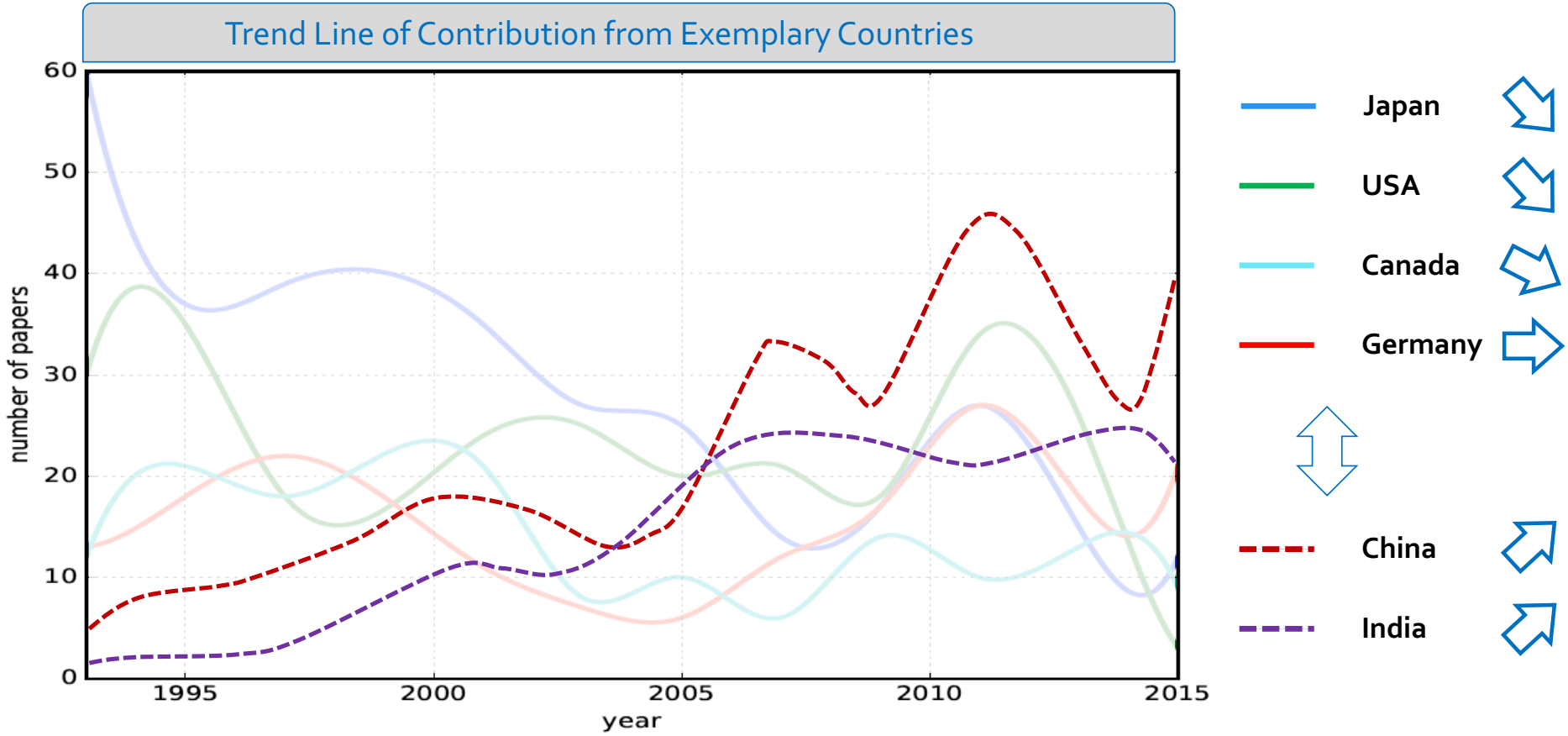
In the period between 1993 and 2015, more than 3,500 authors from 55 countries published papers in the ICDAR proceedings

Macro Indicators (not taken from statistics but from document analysis results)

- The number of accepted papers in 6 years period continuously increased from 634 (1993-1997) to 586 (1999-2003) to 742 (2005-2009) to 789 (2011-2015)*
- The average community member authored 6.01 papers
- The average number of co-authors per paper continuously increased from 3.23 (in 1997) to 4,75 (in 2015)
- The community clustering coefficient is quite high compared to other communities, i.e. authors collaborate more frequently compared to other communities we investigated
- The largest connected component of the ICDAR community co-citation network emphasizes increased from 17% in 1993 to 46% in 2003 to 60% in 2009 and 70% in 2015
- The total number of co-citations between 2005 and 2015 has increased by 150% although the number of publications remain almost the same

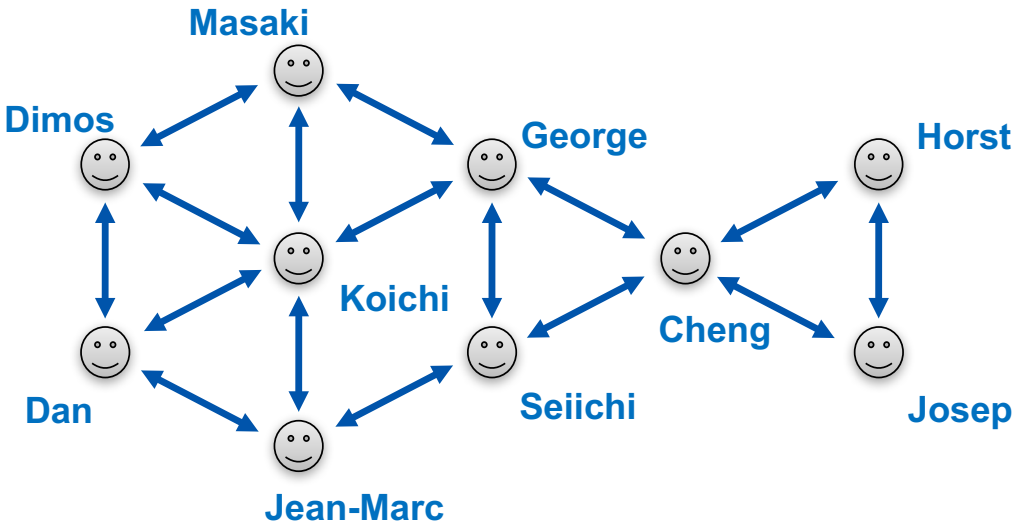
Note: in 2017 we have a total of 212 (52 oral & 160 poster) papers

Regarding the countries contributing to ICDAR conferences there are obvious long term trends



Note: From the early beginning of ICDAR, France has continuously contributed a high number of papers!

More interesting for us is the community network itself



$\mathcal{M}$	Di	Da	M	K	JM	G	S	C	H	J
Dimos	-	1	1	1	0	0	0	0	0	0
Dan	1	-	0	1	1	0	0	0	0	0
Masaki	1	0	-	1	0	1	0	0	0	0
Koichi	1	1	1	-	1	1	0	0	0	0
Jean-Marc	0	1	0	1	-	0	1	0	0	0
George	0	0	1	1	0	-	1	1	0	0
Seiichi	0	0	0	0	1	1	-	1	0	0
Cheng	0	0	0	0	0	1	1	-	1	1
Horst	0	0	0	0	0	0	0	1	-	1
Josep	0	0	0	0	0	0	0	1	1	-

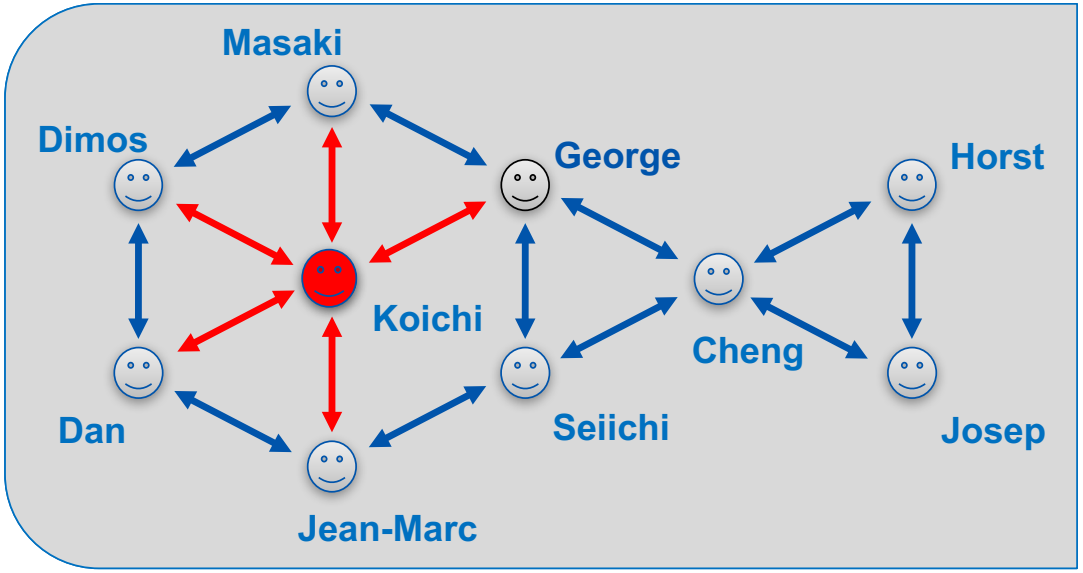
- A co-authorship network is a special kind of a social network where co-authors are represented as nodes n and co-authorship is represented by edges e between two nodes
- As micro indicators in co-authorship and co-citation networks, we **employ centrality measures** helping to find nodes reflecting different dimensions of importance, e.g. influence, leadership, and position

A high degree centrality denotes the existence of authors who collaborate very frequently with other authors

Micro Indicators in co-authorship

⇨ **Degree centrality** DC:= number of ties to directly neighbored nodes

$$C_D(v) = deg(v)$$



DC(Koichi) = 5

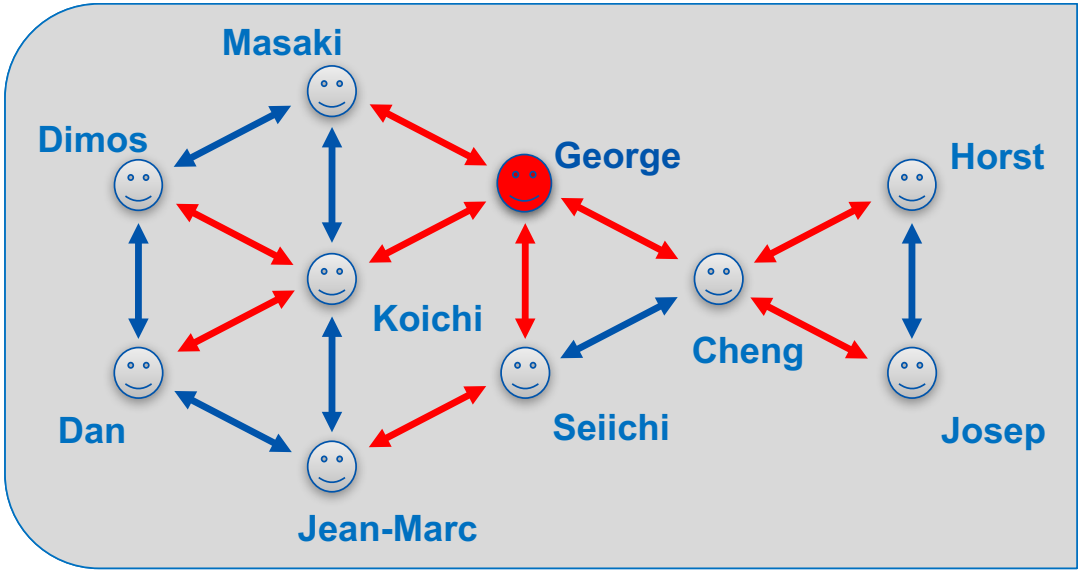
In our case we may distinguish between indegree and outdegree centrality

⇨ Koichi, in our case, is considered as the most central actor

A high closeness centrality means a short path lengths to all other network participants

Closeness centrality CC:= sum of distances (geodesics) to all other nodes

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$



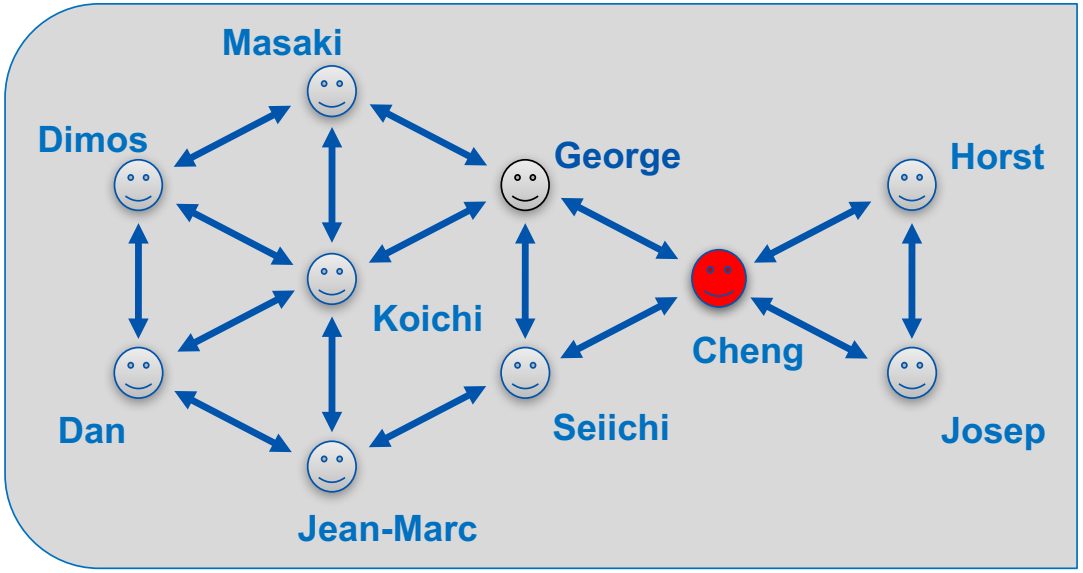
CC(George) = 9/14

George and Seiichi, in our case, have the highest closeness centrality value

Nodes having a high betweenness centrality may control the flow of information among subsets and have therefore a central and powerful position

Betweenness centrality BC:= fraction of all shortest paths passing through this node

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



BC(Cheng) = 14/36

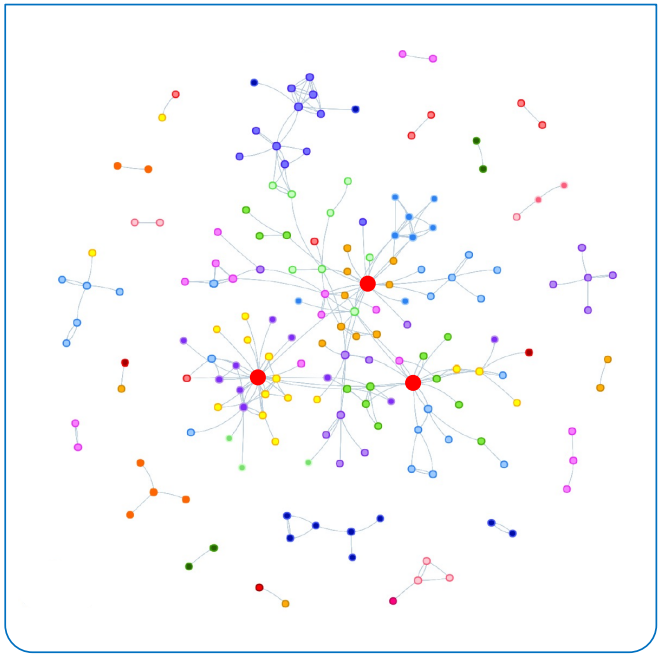
For Cheng this value is 14 (7 geodesics each for Horst and Josep)

Cheng, in our case, has the highest betweenness centrality value

There are some interpretations for the individual centrality measures

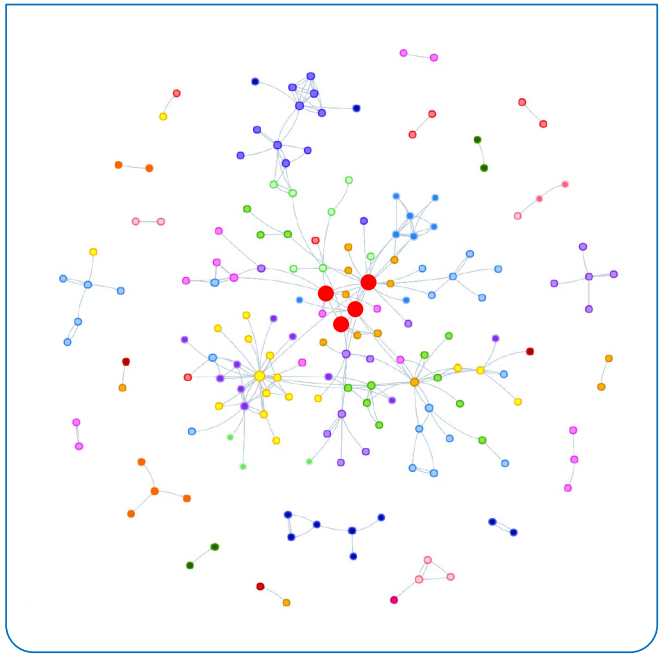
Degree Centrality

Denotes authors who are open and collaborate very frequently with other authors



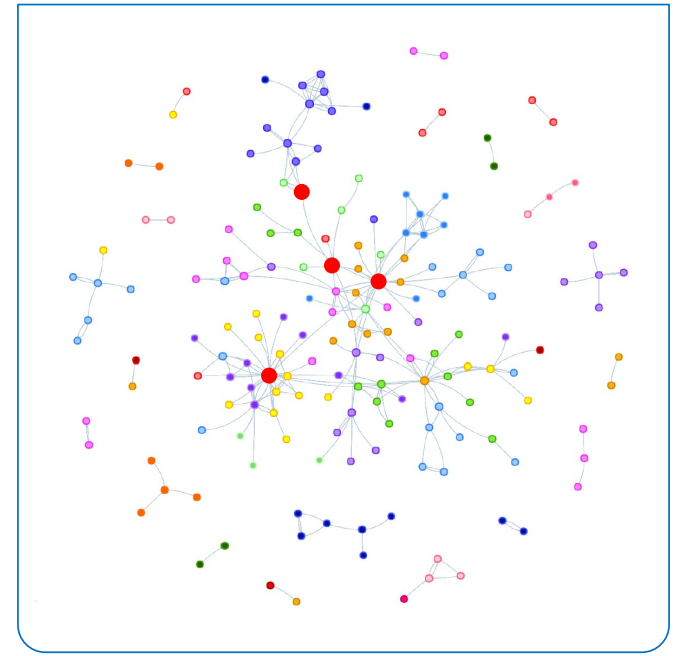
Closeness Centrality

Denotes the prominence an authors as an opinion leader since she/he is well connected

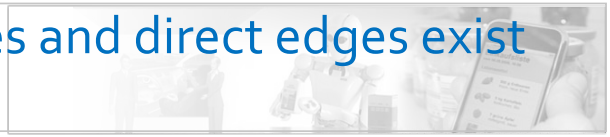


Betweenness Centrality

Denotes authorities who act as bridges between small subgroups in a community (critical ties)



Authors citation networks are networks where authors are nodes and direct edges exist between A and B if "A cites B" in one of his papers



Micro Indicators

- Having a high **in-degree centrality** implies that an author is a dominating person in a community based on her/his scientific contribution (most cited in the case of highest in-degree)
- A high **betweenness centrality** in authors citation networks is given if an author is diversively publishing with different cliques in the community

  In most cases such authors show a almost balanced score in citing others and being cited by others
- An author who has a high **Eigenvector centrality** is considered as being highly influential, e.g. authors who are most likely to produce new research ideas

  An author's Eigenvector centrality is proportional to the sum of the Eigenvector centralities of all nodes directly connected to her/him

In order to find important patterns of co-citation, we cluster author instances into various cliques citing each other's work more often

- Clustering of a social structure is motivated by the principle of `homophily', which is the tendency of individuals to associate and bond with similar others
- Individuals in homophilic relationships share common characteristics (in our case: topics, approaches, applications, , etc.) that make communication and relationship formation easier
- Instead of vertex betweenness, we in this case extend this measure to the case of edges, employing the "edge betweenness", i.e. number of shortest paths between pairs of nodes that run along it
- The **Girvan-Newman Clustering*** algorithm detects communities by progressively removing edges of high betweenness centrality from the original network.

  1. Calculate betweenness centrality for **all** edges existing in the network
  2. Remove edge(s) with highest betweenness centrality
  3. Recalculate betweenness centrality of all remaining edges
  4. Repeat Steps 2 and 3 until no edges are left

* M. Girvan and M. E. Newman. *Community structure in social and biological networks.* Proceedings of the national academy of sciences, 99(12):7821{7826, 2002.

Prominent researchers in ICDAR reveal different roles when considering the normalized centrality values resulting from the analysis

Author	Coauthorship networks			Author citation networks			
	Closeness	Betweenness	Degree	Betweenness	Eigenvector	Outdegree	Indegree
U. Pal	1.0		0.85	0.39	0.18	0.70	0.42
C. Liu				0.65	1.0	1.0	0.80
M. Liwicki			0.72			0.59	0.34
C. Tan	0.93	0.30	0.56			0.57	0.47
S. Uchida	0.92	0.51	0.74			0.44	0.17
C. Suen	0.87	0.52	0.74	0.85	0.08	0.38	0.70
D. Karatzas	0.92	0.29	0.37	0.14			
K. Kise	0.85	0.17	0.45	0.15			
A. Dengel	0.79	0.14	0.49	0.73	0.07		0.33
J. Llados	0.92	0.29	0.55	0.17	0.08	0.41	0.20
S. Srihari	0.83	0.27			0.01	0.17	0.69
M. Nakagawa	0.86	0.25			0.42	0.42	0.30
D. Doermann	0.88	0.21	0.28	0.65	0.07	0.35	0.57
C. Jawahar	0.65	0.21	0.18	0.12	0.06	0.27	0.15
D. Lopresti	0.78	0.06	0.18	0.24	0.03	0.20	0.22
M. Blumestein	0.88		0.26	0.06			0.11
V. Govindaraj				0.49			
M. Cheriet	0.84		0.41	0.36	0.09	0.37	0.26
M. Iwamura	0.87	0.10	0.44	0.04	0.06	0.22	0.14
R. Ingold	0.84	0.22	0.46	0.22	0.06	0.36	0.16

"Opinion Leaders"

"Networkers"

"Community Experts"

"Connectors"

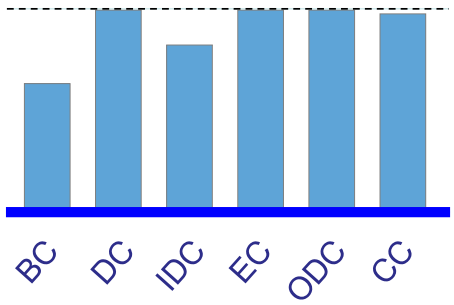
"Collaborators"

"Scientific Leaders"

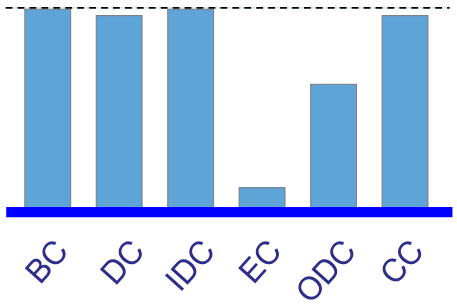
Anyway, the top-influential authors of the ICDAR community show different characteristics, i.e., they play slightly different roles

Results of Quantitative Analysis

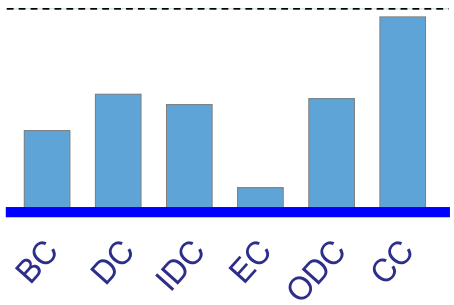
C. LIU	
Betweenness centrality	:0.645475144281
Degree centrality	:1
Indegree centrality	:0.800762631054
Eigenvector centrality	:1
Outdegree centrality	:1
Closeness centrality	:0.996097761509



H. BUNKE	
Betweenness centrality	:1
Degree centrality	:0.962396430875
Indegree centrality	:1
Eigenvector centrality	:0.132500818835
Outdegree centrality	:0.632625085683
Closeness centrality	:0.96363997615



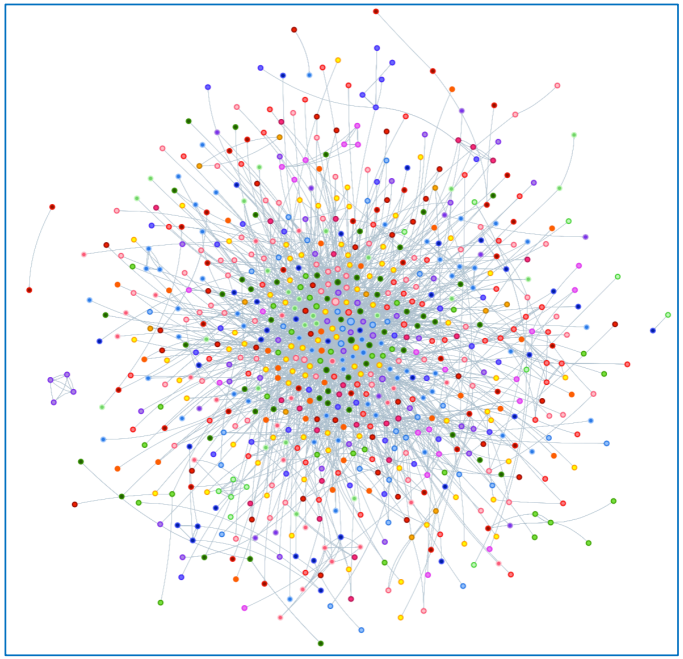
B. GATOS	
Betweenness centrality	:0.340646448672
Degree centrality	:0.622370936879
Indegree centrality	:0.541468064756
Eigenvector centrality	:0.134087160471
Outdegree centrality	:0.560657984943
Closeness centrality	:0.957930080984



For visualizing hidden structures of cliques in the network, we may set different thresholds for the minimum number of co-citations

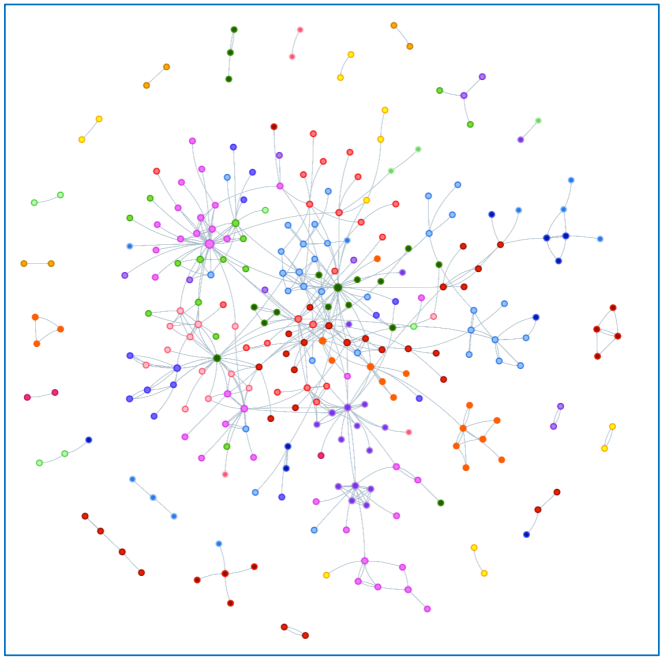
The examples show a cluster or clique from ICDAR community where each edge represents a citation strength of at least *n* between all of its members

The ICDAR Social Network with edges with a minimum citation count of 6!

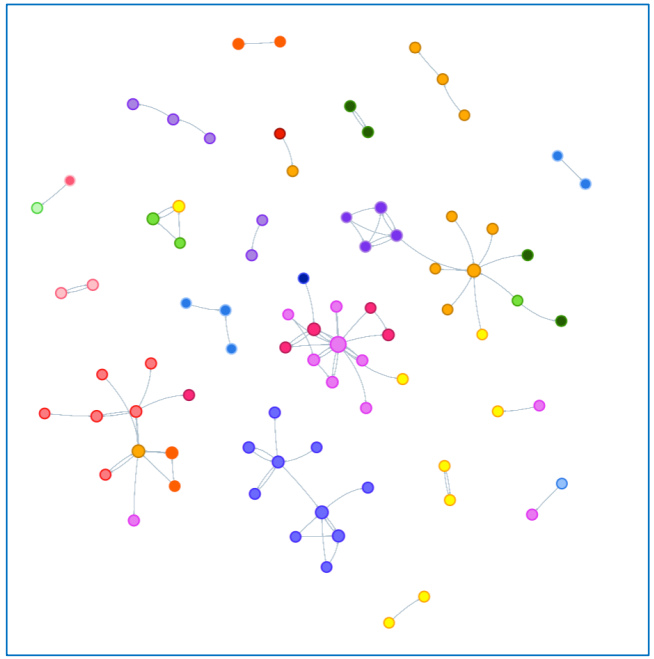


746 authors with 19,239 links

The ICDAR Social Network with edges with a minimum citation count of 11!



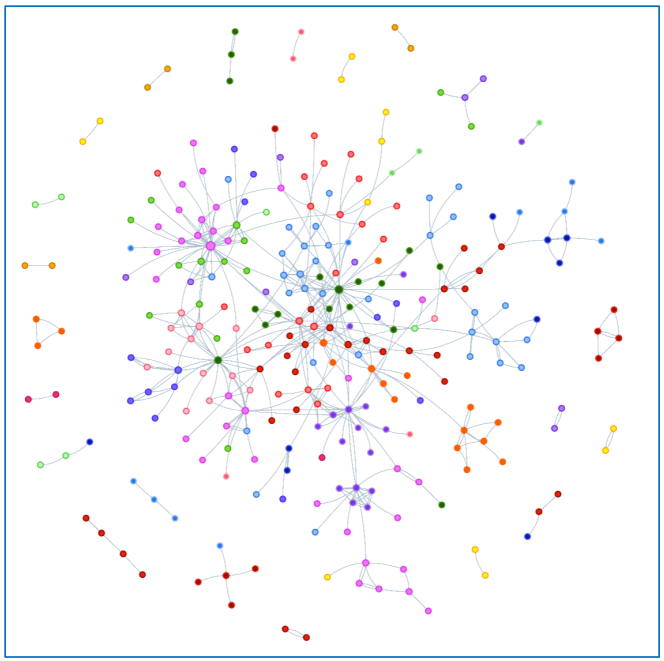
The ICDAR Social Network with edges with a minimum citation count of 21!



81 authors with 2,875 links

Moreover, zooming into the community networks shows the the author connections in more detail where the center captures the most influential authors

The ICDAR Social Network
(Minimum Count of 11) – **Total View** –*



Statistics

Number of Authors :276

Links :8192

Connected Groups

25

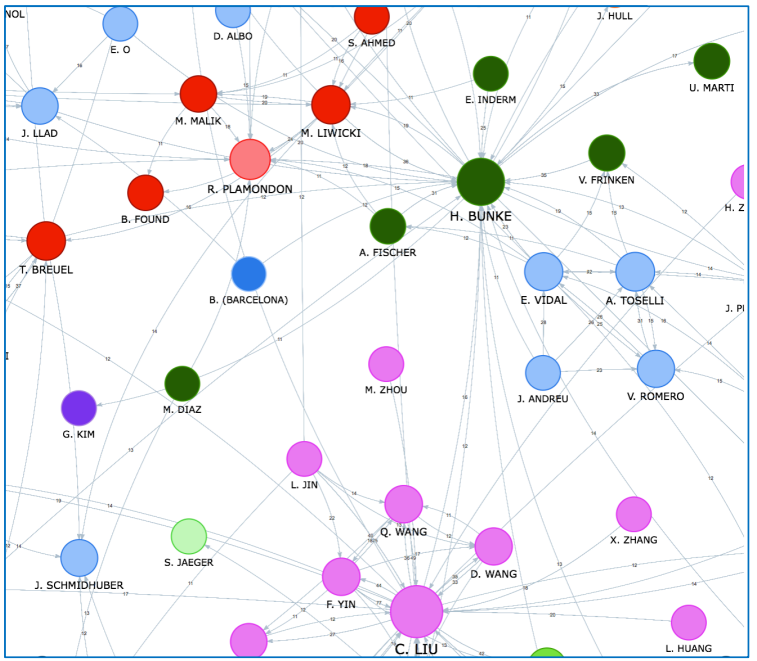
Giant Component has

253 Authors

Most Influential(s) in Giant Component

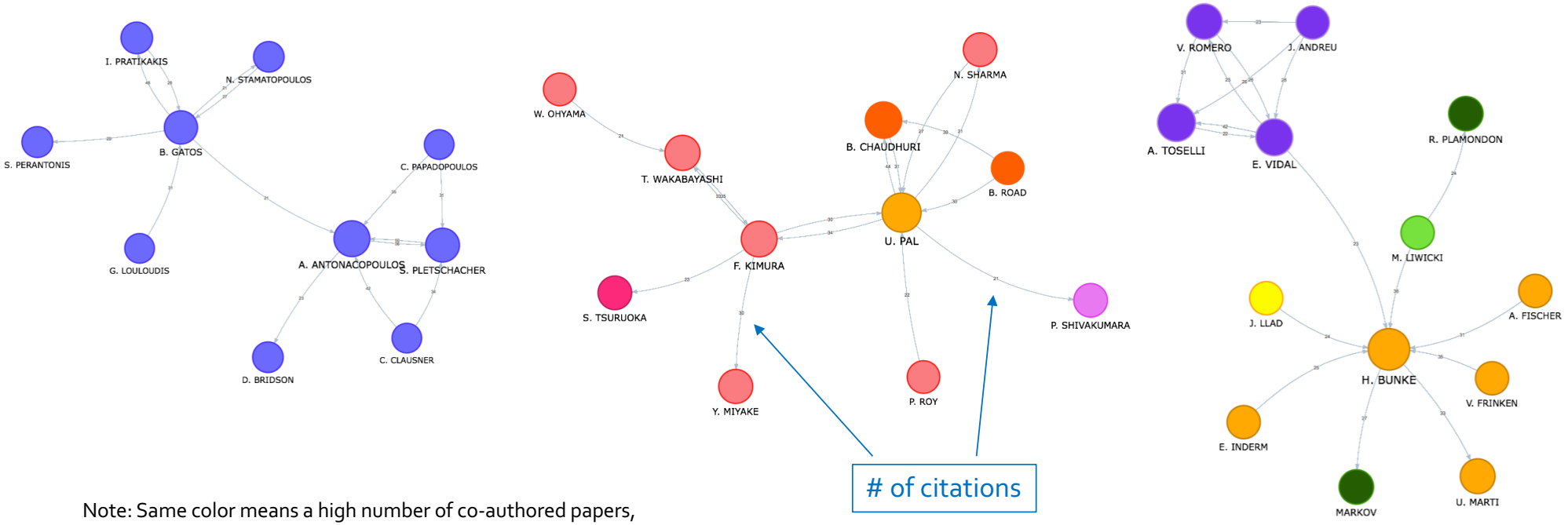
C. Liu

The ICDAR Social Network with edges
(Minimum Count of 11) – **Center Zoom** –



Having a closer look, we receive cliques of authors who collaborate and cite each other frequently

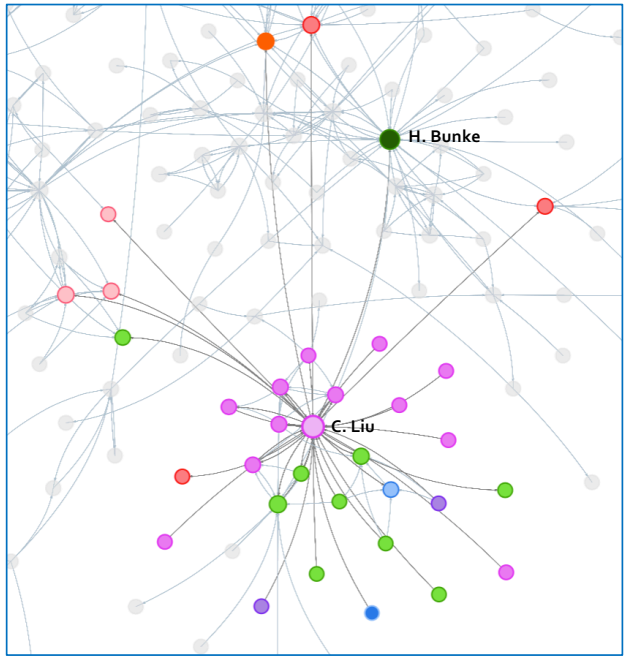
The examples shows clusters or clique from ICDAR community where each edge represents a citation strength of at least 21 between all of its members



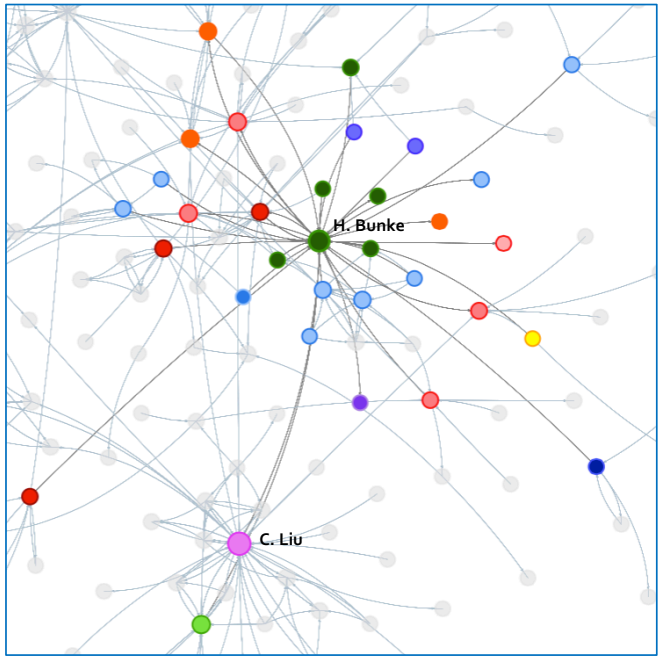
Note: Same color means a high number of co-authored papers, i.e. intensive collaboration

Clicking on a node shows individual citation networks of authors, i.e most influenced community members

The ICDAR Social Network of **C. Liu**
(Minimum Citation Count of 11)



The ICDAR Social Network of **H. Bunke**
(Minimum Citation Count of 11)

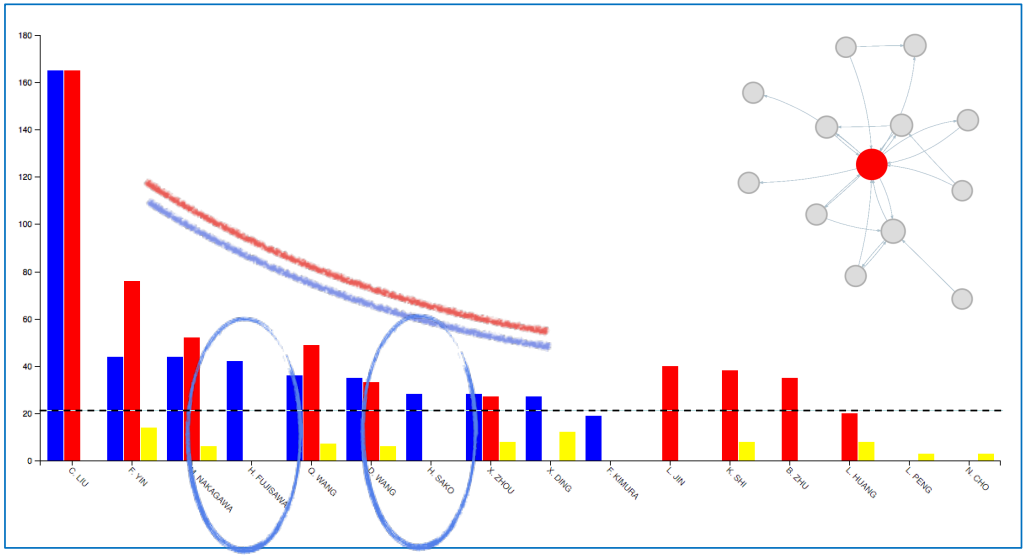


Statistics

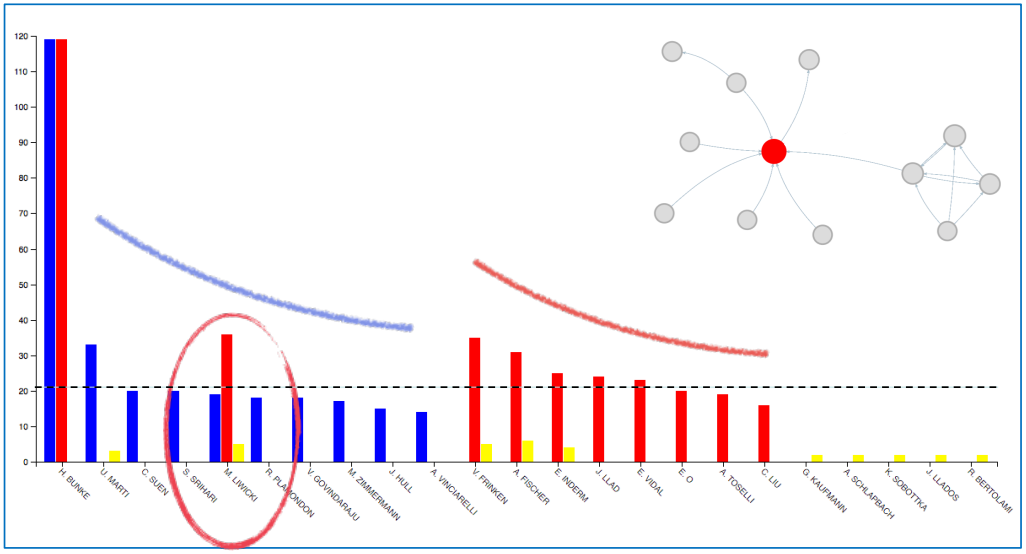
C. LIU(537)
H. BUNKE(387)
U. PAL(217)
B. GATOS(194)
M. NAKAGAWA(192)
R. PLAMONDON(170)
A. ANTONACOPOULOS(167)
C. TAN(163)
C. SUEN(158)
S. SRIHARI(153)
B. CHAUDHURI(151)
F. KIMURA(147)
I. PRATIKAKIS(129)
V. GOVINDARAJU(128)
A. TOSELLI(127)
H. FUJISAWA(119)
T. BREUEL(118)
M. LIWICKI(114)
E. VIDAL(106)
P. SHIVAKUMARA(104)

Different authors show different data characteristics reflected by the **Overlap Index Graph** aggregating and showing the diversity of an author in the community

Citation Behavior in the context of **C. Liu**
(Minimum Count of 21)



Citation Behavior in the context of **H. Bunke**
(Minimum Count of 21)



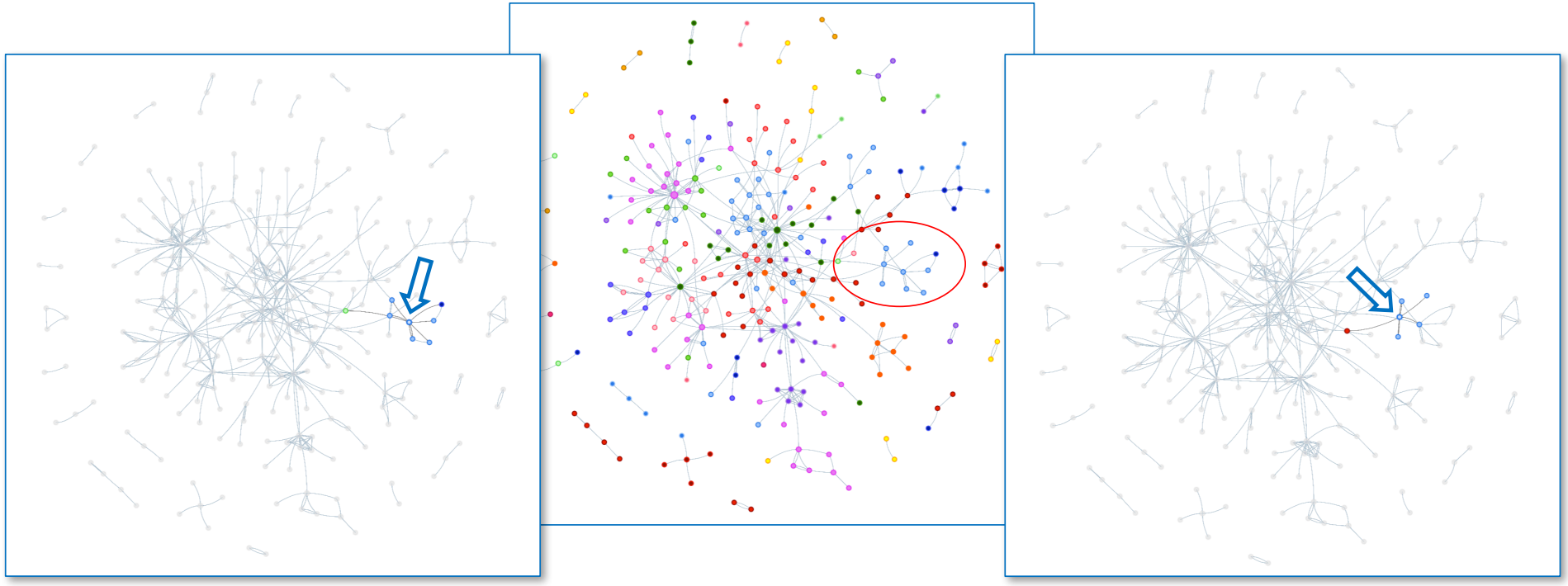
cites most often

most often cited by

collaborates most often with

The graph may be investigated via navigation options

The ICDAR Social Network with edges with a minimum citation count of 11!



When clicking on one node in the network, we get a detailed overview of all relevant data dimensions

J. OGIER

Cites most often

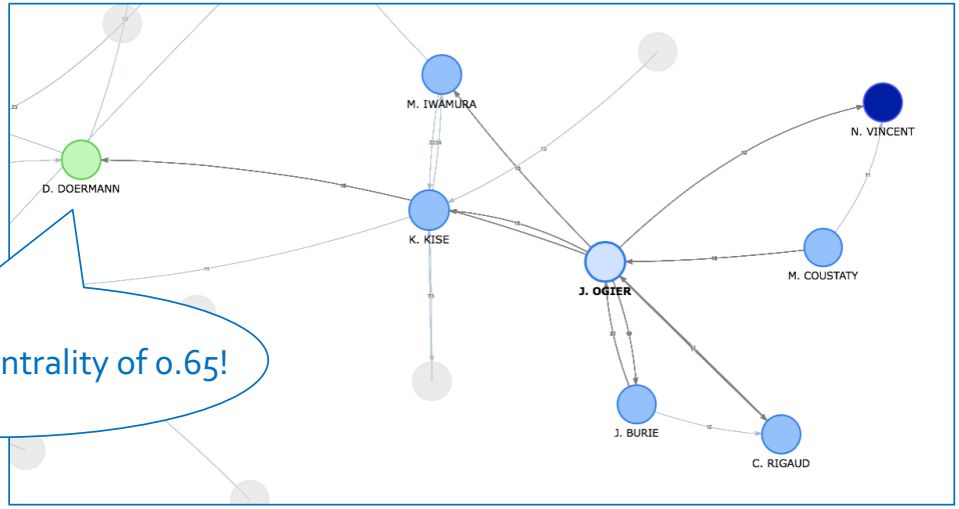
Count	Author name
39	J. OGIER
19	J. BURIE
18	D. DOERMANN
15	K. KISE
13	M. IWAMURA
12	N. VINCENT
11	C. RIGAUD
10	M. RUSINOL
10	M. COUSTATY
9	J. LLAD

Most often cited by

Count	Author name
39	J. OGIER
21	J. BURIE
18	M. COUSTATY
11	C. RIGAUD
10	M. LUQMAN
8	M. MEHRI
7	S. ROUVRAY
7	P. H
7	P. AMER
6	K. KISE

Collaborates most often with

Count	Author name
9	J. BURIE
6	M. COUSTATY
5	M. LUQMAN
5	J. CHAZALON
4	C. TRAN
4	M. RUSINOL
4	N. NAYEF
4	S. PRUM
3	C. RIGAUD
3	M. VISANI



Betweenness Centrality of 0.65!

J. OGIER

Betweenness centrality	:0.215982390334
Degree centrality	:0.332058636116
Indegree centrality	:0.183508103029
Eigenvector centrality	:0.0713975993674
Outdegree centrality	:0.450993831373
Closeness centrality	:0.968257137334
Positive references	:25.55%
Neutral references	:63.50%
Negative references	:10.95%
List of all papers	01) A Pixel Labeling Approach for Historical Digiti... 02) A Toplogy Based Multi-Classifier System 03) A method for image local-difference visualization 04) Automatic Annotation Extension and Classificatio... 05) Knowledge-based recognition of utility map sub-... 06) Layout Analysis for Historical Manuscripts Usin... 07) Statistical Modeling of the Relation Between Ch... 08) Text/Graphics Segmentation in Architectural Flo... 09) Word-Based Adaptive OCR for Historical Books * 10) eBDtheque: a representative database of comics

When clicking on one node in the network, we get a detailed overview of all relevant data dimensions

K. KISE

Cites most often

Count	Author name
45	K. KISE
24	M. IWAMURA
13	K. KUNZE
11	A. DENGEL
9	S. UCHIDA
8	S. OMACHI
7	H. KAWAICHI
7	K. YOSHIMURA
6	J. OGIER
6	T. NAKAI

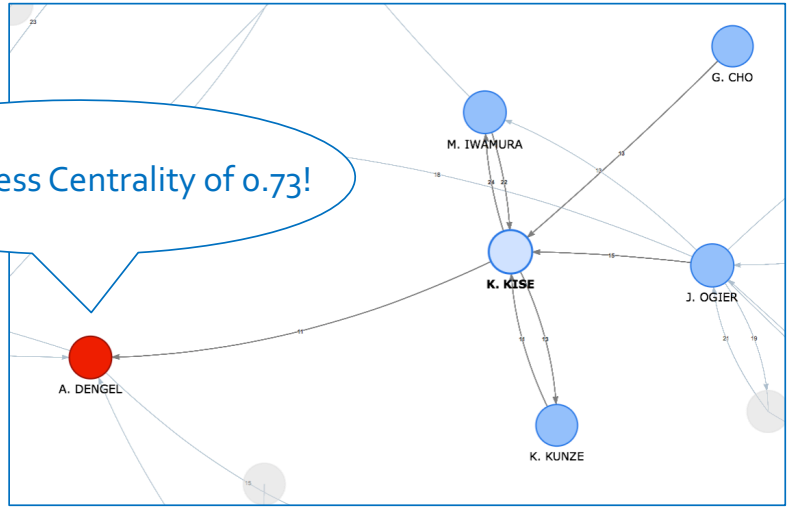
Most often cited by

Count	Author name
45	K. KISE
22	M. IWAMURA
15	J. OGIER
13	G. CHO
11	K. KUNZE
10	C. TRAN
10	M. COUSTATY
10	M. LUQMAN
10	Q. DANG
8	B. (BARCELONA)

Collaborates most often with

Count	Author name
11	M. IWAMURA
5	G. CHO
4	K. MATSUMOTO
3	A. DENGEL
3	K. KUNZE
3	T. NAKAI
2	K. YOSHIMURA
2	S. OMACHI
2	S. UCHIDA
2	T. KOBAYASHI

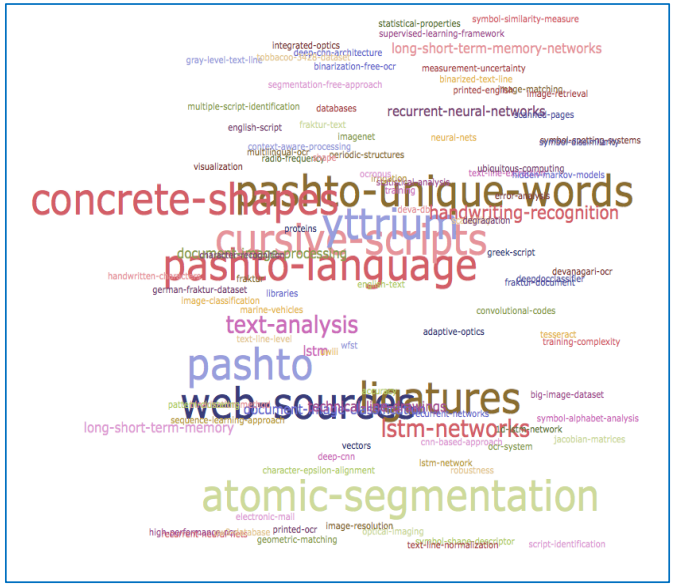
Betweenness Centrality of 0.73!



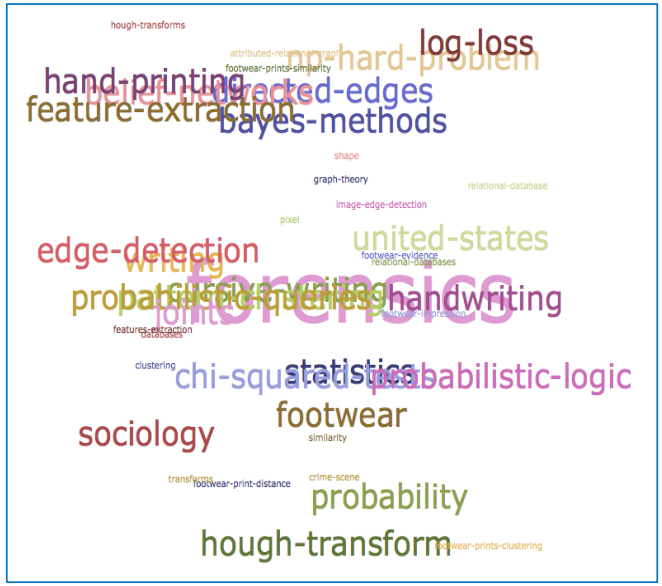
K. KISE	
Betweenness centrality	:0.150754429771
Degree centrality	:0.239005736174
Indegree centrality	:0.203050524216
Eigenvector centrality	:0.0424678780914
Outdegree centrality	:0.222755311853
Closeness centrality	:0.880217785945
Positive references	:40.37%
Neutral references	:47.71%
Negative references	:11.93%
List of all papers	01) Automatic Ground Truth Generation of Camera Cap... 02) Construction of Generic Models of Document Stru... 03) Detection of Cut-And-Paste in Document Images 04) Document Understanding System Using Stochastic ... 05) Extending Page Segmentation Algorithms for Mixe... 06) Graphics Recognition – from Re-engineering to R... 07) Real-Time Camera-Based Recognition of Character... 08) Recognizing Characters with Severe Perspective ... 09) Trajectory Recovery and Stroke Reconstruction o... 10) Wearable Reading Assist System: Augmented Reali...

Furthermore, topic clouds offer insights into an author's domain of research at a glance

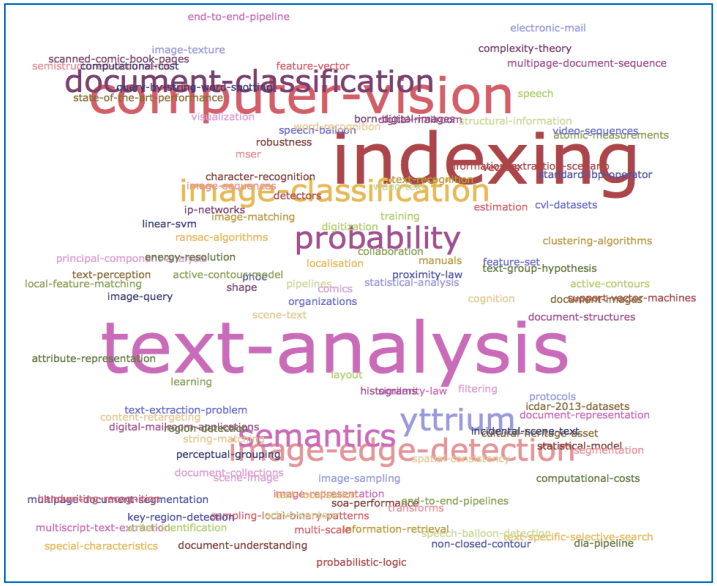
Topic Cloud of **T. Breuel**
(Minimum Count of 11)



Topic Cloud of **S. Srihari**
(Minimum Count of 11)



Topic Cloud of **D. Karatzas**
(Minimum Count of 11)



In addition, we try to address another very challenging problem which is of high relevance but very hard to solve (work in progress)

- Motivation:
  - There are behavioral patterns of researchers' literature review practice in each community capturing the author's opinion toward the cited work
  - Citation Sentiment Analysis (CSA) is a newly emerged research topic inspired by traditional citation context analysis in scientometrics and applied linguistics
  - In a so-called **Senti-Index**, we intend to consider citation sentiment as the polarity of a citing author's opinion using the categories positive, negative and neutral (or many more)

- Just to show (again) the target of other communities:
  - **Document Analysis** is a form of qualitative research in which documents such as public records of an organization (community) are interpreted by the researcher to give voice and meaning around an assessment topic, e.g. to provide a confluence of evidence that breeds credibility

While for the SNA we focused on authors, abstract, keywords and list of references, in CSA we analyzed the text of the paper including the captured references

Text

Preprocessing

1. Tokenizer
2. Regular Expression Sentence Splitter

Tagged Structured Text

Filtering

1. Reference Transducer
2. Reference Sentence Transducer
3. TokenML Transducer

Tagged Struct. Ref. Sentences

Information Extraction

1. POS Tagger
2. WorNet Suggester
3. Stemmer

Features

Machine Learning

1. Sentiment Analysis
2. Nature of References

Reference Polarity Score

where N denotes a neighborhood system (we use 8-connectedneighborhood in experiments), x denotes pixel coordinates, cmeans RGB color, σg and σc are normalization constants, λdetermines the degree of smoothness. The pairwise term thusimposes a cost for the boundaries in the binarization resultaccording to the local color contrast in the input image.

	DS-I was used to evaluate the Performance [12].										
Root	Ds-I	Be	Use	To	Evaluate	The	Performance	[	12	]	.
Orth	allCaps	Lower case	Lower case	Lower case	Lowercase	Lower case	Lowercase	-	-	-	-
Category	NNP	VBD	VBN	TO	VB	DT	NN	NN	CD	NN	.
Kind	Word	Word	Word	Word	Word	Word	Word	Punct ation	Number	Punct ation	Punct ation
String	DS-I	Was	Used	To	Evaluate	The	Performance	[	12	]	.
Stem	ds-i	Was	Use	To	Evalu	The	Perform	[	12	]	.
Synonyms	-	-	-	-	Measure, evaluate, valuate, assess, appraise, value	-	Performance, public_presen tation	-	-	-	-
Hypernyms	-	-	-	-	Evaluate, pass_judgm ent, judge	-	show	-	-	-	-

List of tags with corresponding part of speech

/IN	preposition or subordinating conjunction
/DT	determiner
/JJ	adjective
/PRP$	possessive pronoun
/NN	common noun, singular
/NNP	proper noun, singular
/NNS	common noun, plural
/VBN	verb, past participle
/VBZ	Verb, 3rd person singular present
/.	
/,	

A. Bhardwaj, D. Mercier, S. Ahmed, and A. Dengel, *DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction*, ICONIP 2017, 24th Int'l Conference on Neural Information Processing, Gouangzhou, China, (Nov. 2017)

D. Mercier, A. Bhardwaj, S. Ahmed, and A. Dengel, *SentiCite: An Approach for Publication Sentiment Analysis*, ICAART-18, 10th Int'l Conference on Agents and Artificial Intelligence, Madeira, Portugal (Jan. 2018),

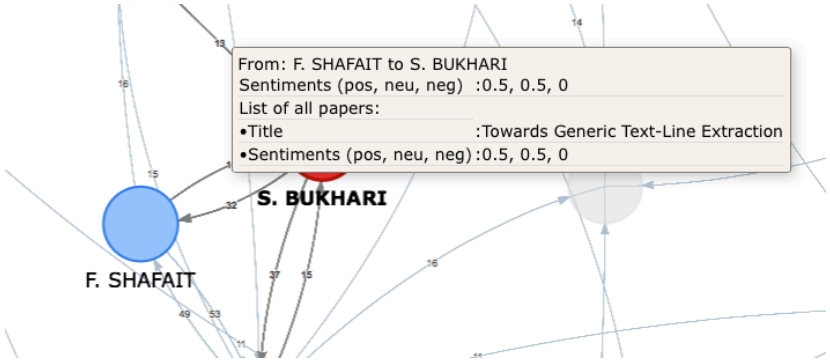
CiTO*, the Citation Typing Ontology, is an ontology to enable characterization of the nature or type of citations, both factually and rhetorically

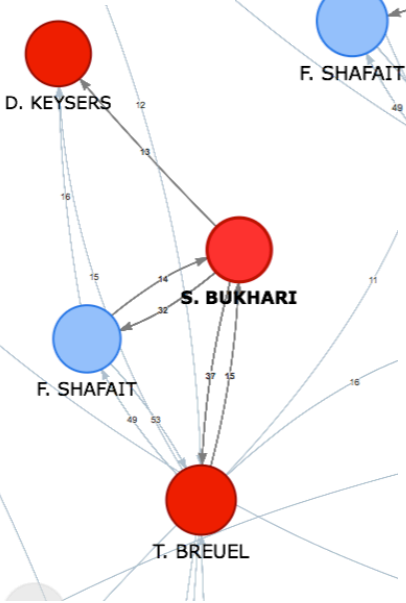
- The citations characterized may be:
  - direct and explicit (as in the reference list of a journal article),
  - indirect (e.g. a citation to a more recent paper by the same research group on the same topic)
  - implicit (e.g. as in artistic quotations or parodies, or in cases of plagiarism)

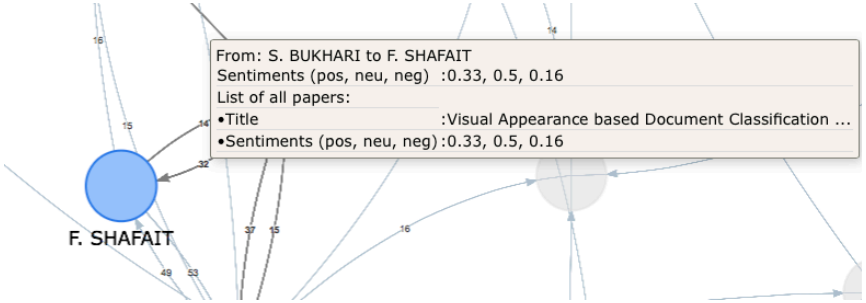
- Properties: agreesWith, cites, citesAsAuthority, citesAsDataSource, citesAsEvidence, citesAsMetadataDocument, citesAsRelated, citesAsSourceDocument, citesForInformation, confirms, containsAssertionFrom, corrects, credits, critiques, disagreesWith, discusses, disputes, documents, extends, givesBackgroundTo, givesSupportTo, includesExcerptFrom, includesQuotationFrom, isAgreedWithBy, isCitedAsAuthorityBy, isCitedAsDataSourceBy, isCitedAsEvidenceBy, isCitedAsMetadataDocumentBy, isCitedAsRelatedBy, isCitedAsSourceDocumentBy, isCitedBy, isCitedForInformationBy, isConfirmedBy, isCorrectedBy, isCreditedBy, isCritiquedBy, isDisagreedWithBy, isDiscussedBy, isDisputedBy, isDocumentedBy, isExtendedBy, isParodiedBy, isPlagiarizedBy, isQualifiedBy, isRefutedBy, isReviewedBy, isRidiculedBy, isSupportedBy, isUpdatedBy, obtainsBackgroundFrom, obtainsSupportFrom, parodies, plagiarizes, providesAssertionFor, providesDataFor, providesExcerptFor, providesMethodFor, providesQuotationFor, qualifies, refutes, reviews, ridicules, sharesAuthorsWith, supports, updates, usesDataFrom, usesMethodIn,

* D. Shotton and S. Peroni, *CiTo: The Citation Typing Ontology*, http://www.sparontologies.net/ontologies/cito/source.html

The Senti-Index allows to hover over a node or edge and get an idea about an authors (author2author) citation sentiment







Take-Aways!

- I presented an approach for an holistic analysis of an entire document corpus which in combination of all containing documents represents a rich source of information
- All results shown are available via the Academic Community Explorer framework (cf. http://www.dfki.uni-kl.de/ace/) analyzing scholarly document metadata to study a scientific community
- The topic clouds offer an overview of an author's domain of research and may be used to search for appropriate reviewers for ICDAR and IJDAR
- The employment of the co-authorship network may avoid "conflict of interest" cases when assigning reviewers because it implicitly shows frequent co-authorship
- The Senti-Index is a first but promising approach to disclose sentiment in citations but has to be considered with great caution
- Anyway, finally this work will continue and should be used as a recommending and exploring system for community members

Finally, I hope I was able to stimulate discussion regarding both, document analysis research questions/options and the hidden patterns of our community

Special thanks to
Akansha Bhardwaj, Sheraz Ahmed,
Dominique Mercier, Hisham Hashmi

Prof. Andreas Dengel
DFKI GmbH
P.O. Box 2080
D-67608 Kaiserslautern
email: andreas.dengel@dfki.de

Questions?