Characterizing a social bookmarking and tagging network

Evangelos Milios

Dalhousie Univ., Faculty of Computer Science

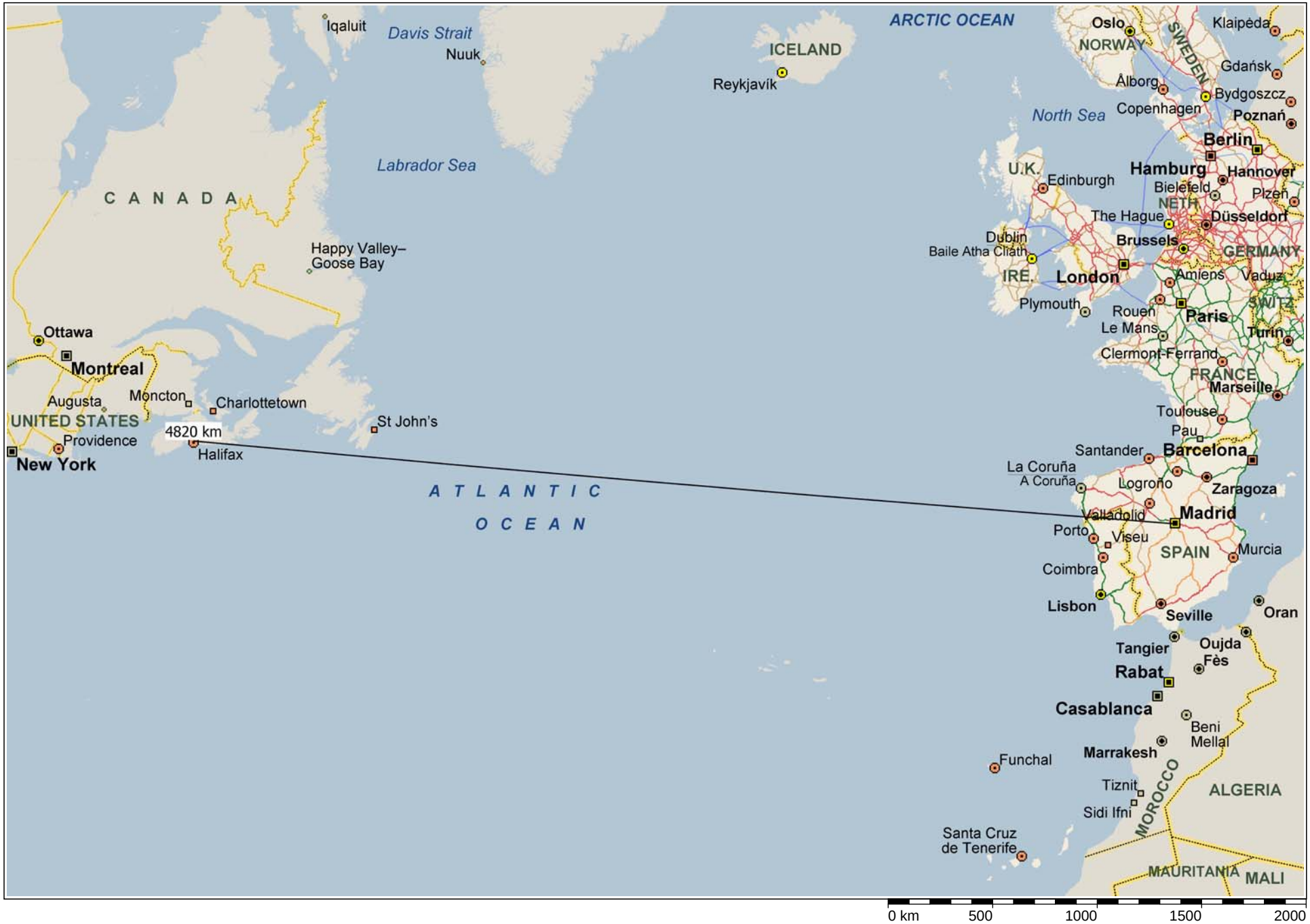
Joint research with
Ralitsa Angelova[1]
Marek Lipczak[2],
Paweł Prałat[2]

[1] Max Planck Institut für Informatik, Saarbrücken, Germany

[2] Dalhousie University

DALHOUSIE UNIVERSITY
Inspiring Minds

World



Bird’s eye view of Halifax

Halifax Fun

Halifax, Nova Scotia

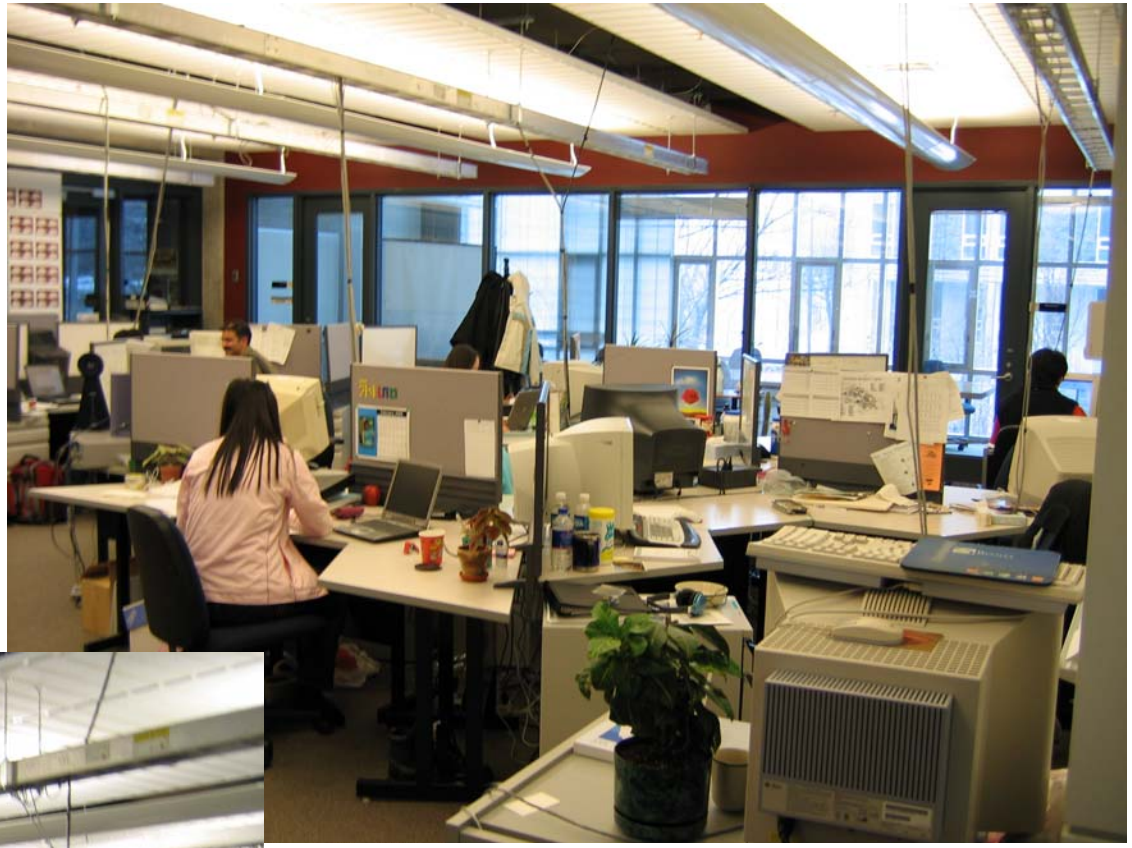
- Northernmost harbour that does not freeze in the winter
- Relatively mild climate
- Metropolis of Atlantic Canada (incl. Nova Scotia, New Brunswick, Prince Edward Island, and Newfoundland)
- Regional economic, cultural and research hub
- Settled by Europeans in the 18th century

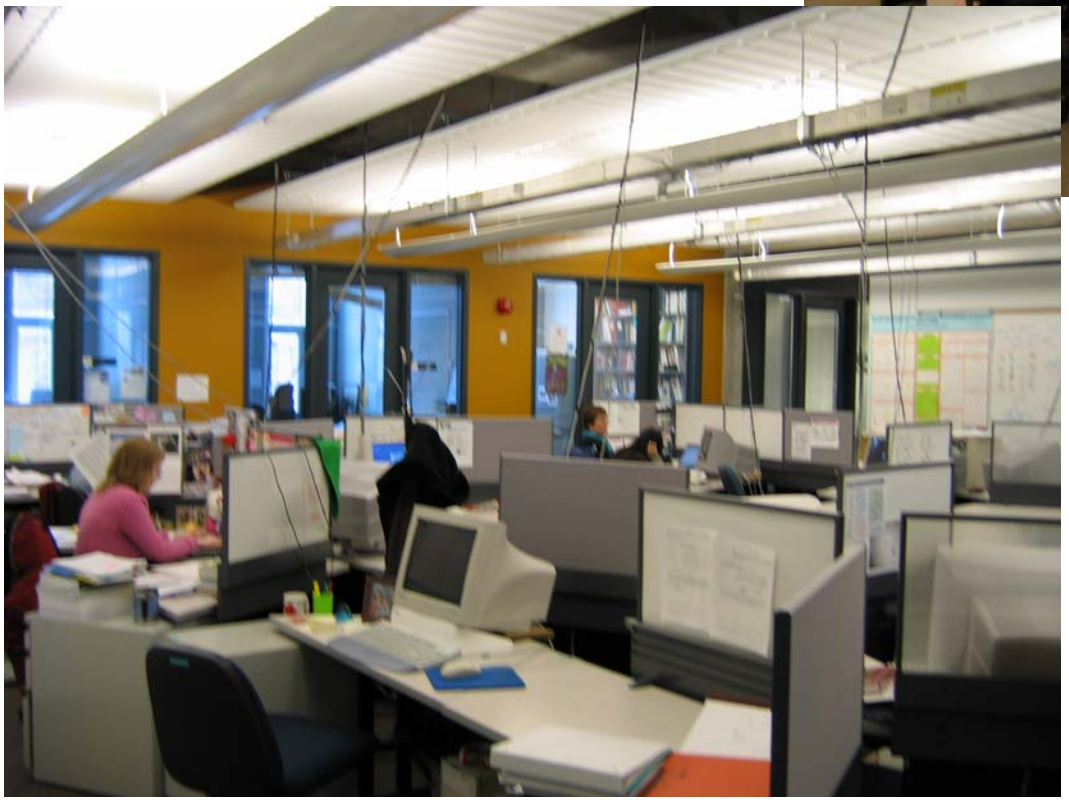
Dalhousie U. Facts

- Founded in 1818
- The smallest Medical/Doctoral university in Canada
  - Medical school
  - Law school
  - Engineering
  - Business school
- World class
  - Oceanography
  - Biology
  - Medicine
  - Sciences
- Regional Research Hub for Atlantic Canada

Faculty of Computer Science


SECOND CUP





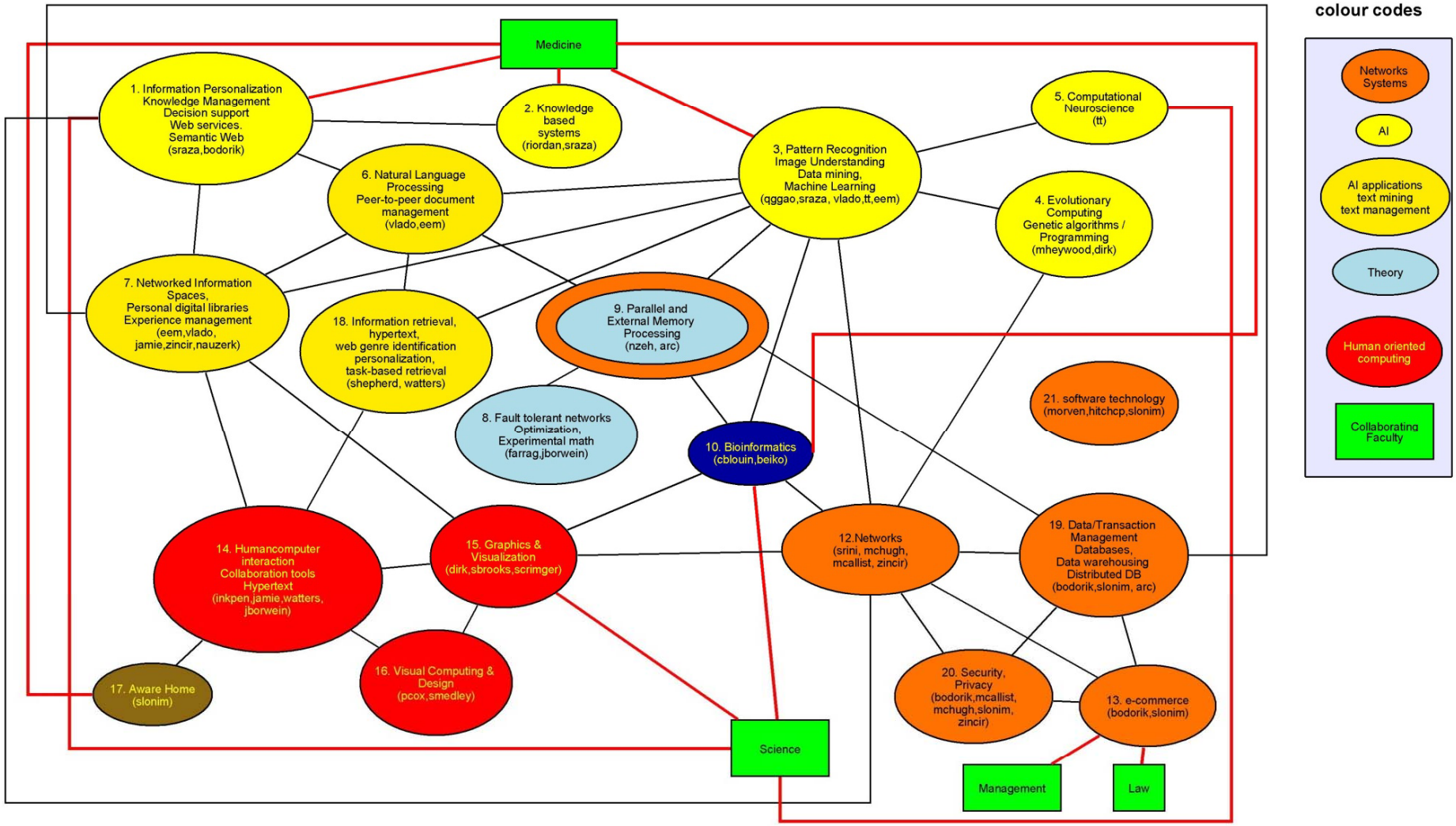
Faculty of Computer Science

- Established in 1997
- Strengths in:
  - Information retrieval, text mining
  - Health informatics & Knowledge management
  - Bioinformatics
  - Human-computer interaction
  - Computer networks, network management, intrusion detection
  - Algorithms, graph theory, parallel computation

Interdisciplinary outlook

- Master’s degrees in:
  - Computer Science
  - Health informatics (with Medicine)
  - Electronic commerce (with Business and Law)
  - Bioinformatics (with Biology)
- Joint research projects with
  - Mathematics
  - Engineering
  - Medicine
  - Business
  - Biology

Research Profile of the Faculty



Research snippets

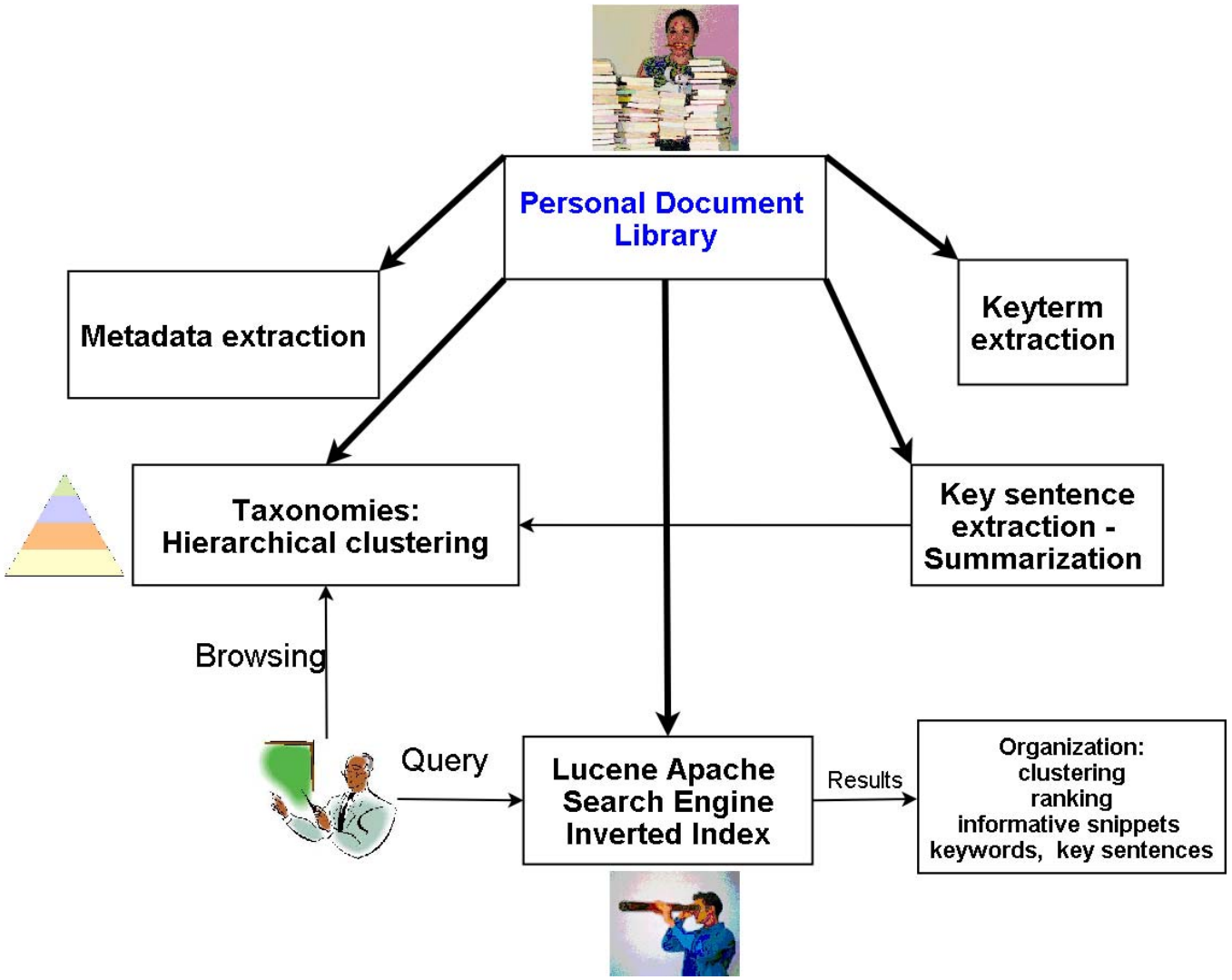
Networked Information Spaces:

Modelling and Mining

Documents are networked into information spaces

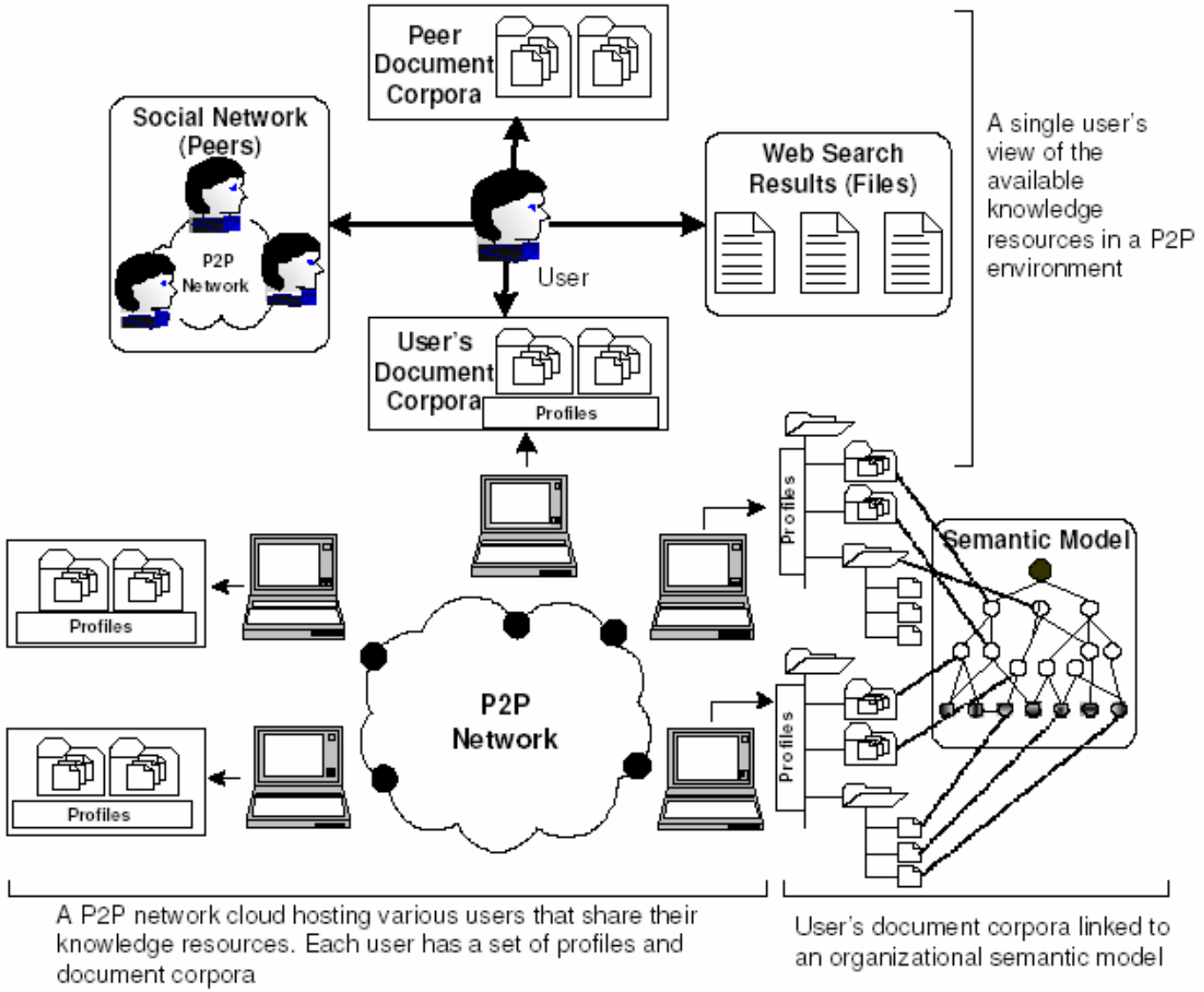
- World Wide Web
- Blog space
- Scientific and Medical Literature
- Patents
- Common Law

Desktop of the future



Peer-to-Peer Document Management

V. Keselj, E. Milios, S. Abidi



Automatic Topic Extraction

E. Milios

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9
error	neuron	image	analog	data	control	function	rule	distribution
generalization	neurons	images	circuit	clustering	model	functions	rules	probability
learning	synaptic	object	current	principal	motor	basis	set	gaussian
training	firing	recognition	figure	cluster	forward	linear	step	data
optimal	spike	face	chip	pca	inverse	regression	form	parameters
order	time	objects	voltage	set	dynamics	kernel	fuzzy	model
large	activity	hand	vlsi	algorithm	controller	space	problem	bayesian
average	rate	pixel	circuits	points	feedback	gaussian	relative	mixture
small	synapses	system	digital	approach	system	approximation	extraction	density
examples	potential	view	implementation	clusters	position	rbf	expert	likelihood

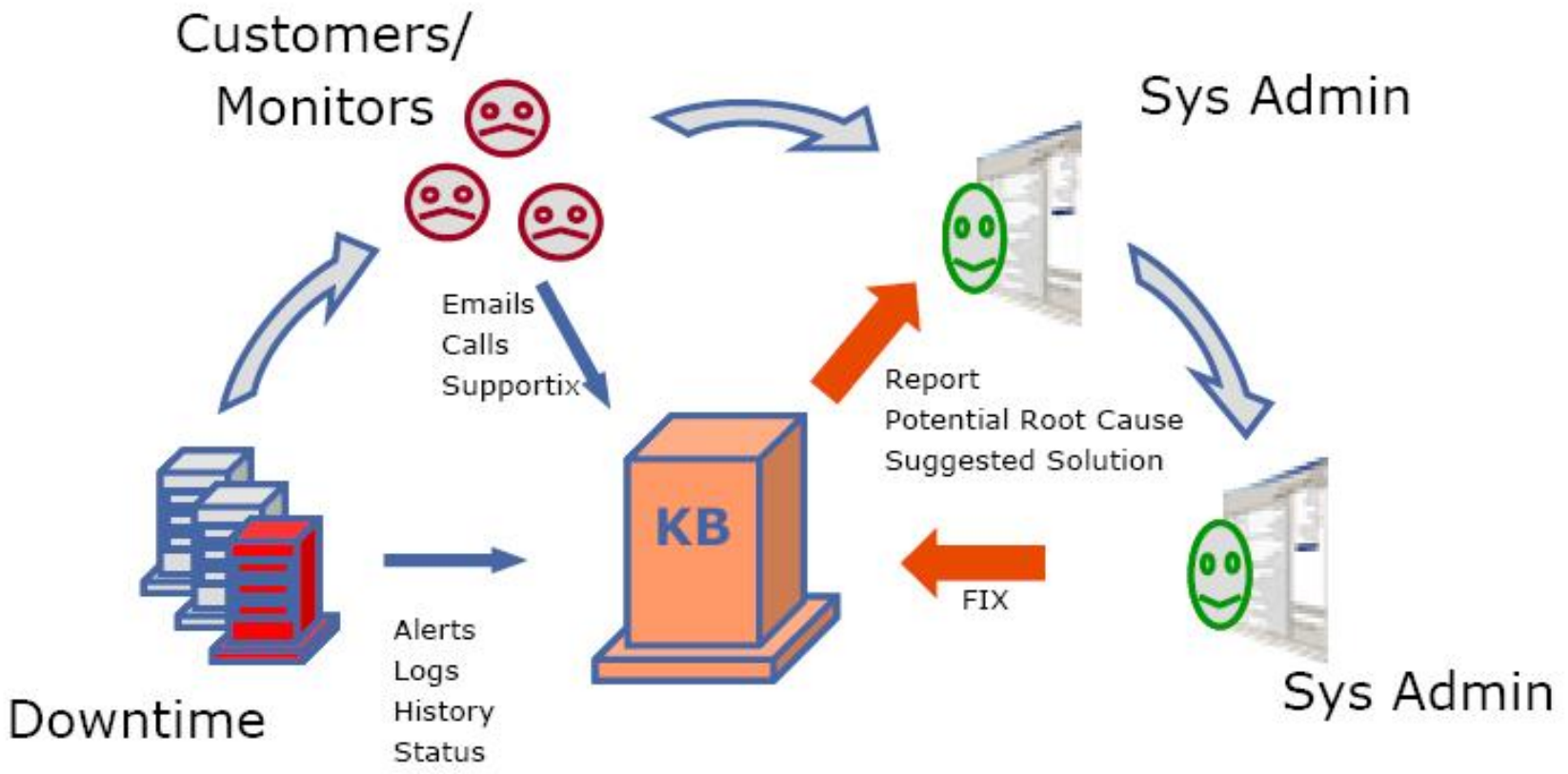
Figure 2. Example word-topics for the NIPS dataset

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
language	game	church	house	air	league	war	apollo	party	system
english	player	god	parliament	aircraft	football	german	earth	government	computer
greek	cards	christian	members	world	team	army	moon	president	game
languages	players	jesus	commons	force	world	soviet	lunar	political	games
word	games	christ	lords	military	club	battle	time	national	apple
russell	play	orthodox	bill	ship	home	germany	mission	minister	atari
century	card	baptism	act	gun	season	world	program	states	commodore
theory	hand	life	power	war	won	forces	module	united	home
words	round	catholic	chopin	ships	game	french	jpg	election	software
modern	played	roman	speaker	navy	major	union	crew	state	video

Figure 3. Example word-topics for the Wikipedia dataset

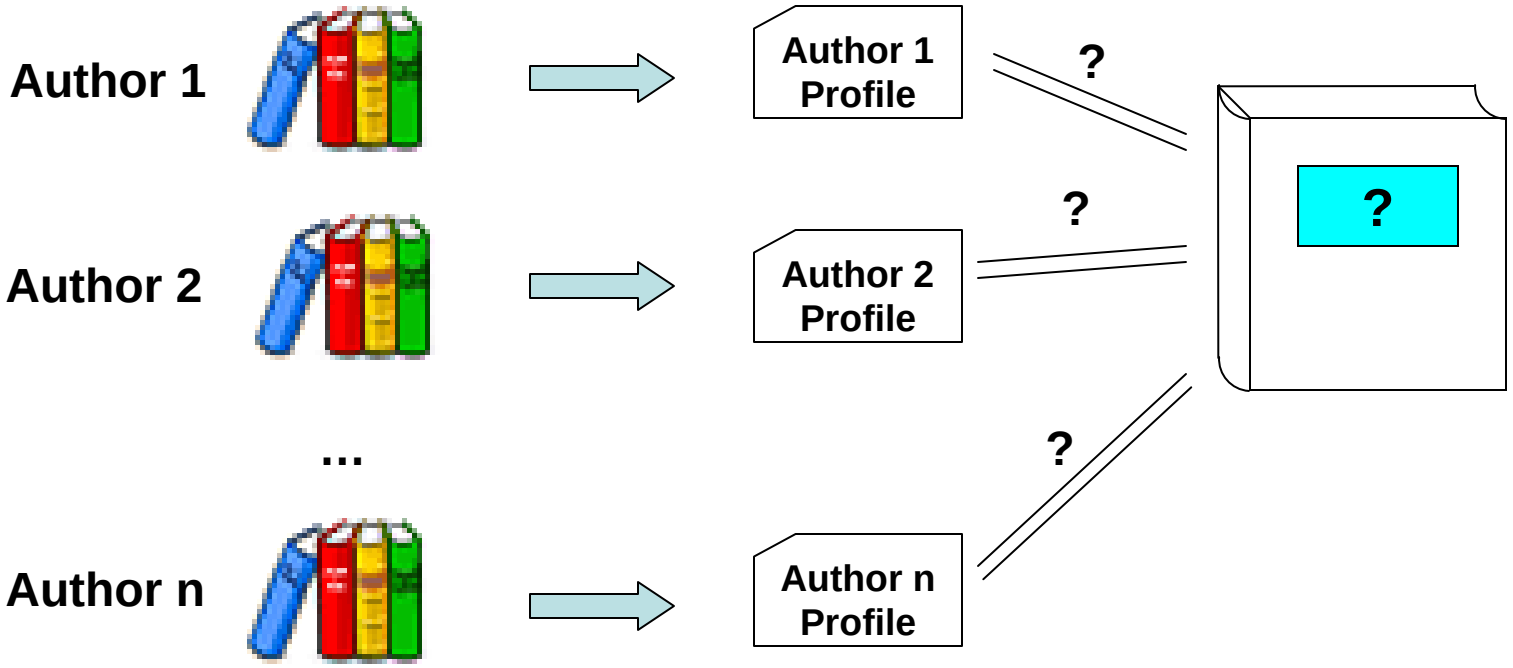
Experience Management

E. Milios, N. Zincir-Heywood



Authorship Attribution using Character N-grams

Vlado Keselj

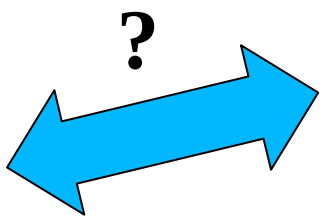


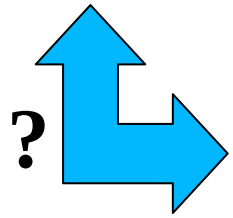
Dickens: A Tale of Two Cities

```
_th 0.016
the 0.014
he_ 0.012
and 0.007
nd_ 0.007
```

Dickens: Christmas Carol

```
_th 0.015
___ 0.013
the 0.013
he_ 0.011
and 0.007
```



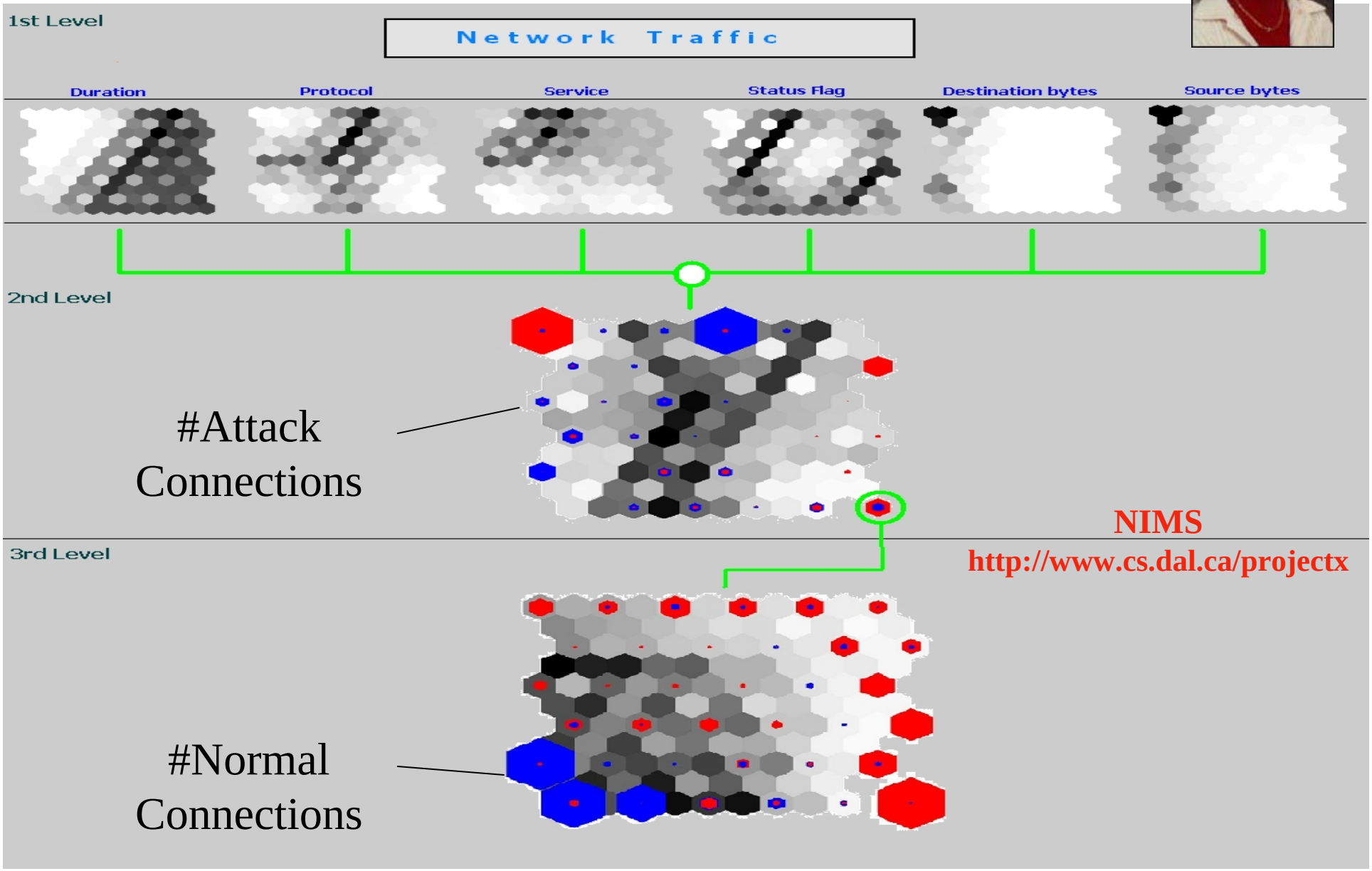


Carroll: Alice's adventures in wonderland

```
_th 0.017
___ 0.017
the 0.014
he_ 0.014
ing 0.007
```

Network Traffic Classification

Nur Zincir-Heywood



NIMS

http://www.cs.dal.ca/projectx

The Graphics and Visualization Lab

- The focus is on both:
  - the development of new graphical techniques, and
  - the application of those techniques, often in cross-disciplinary areas

- Our lab incorporates expertise in areas such as:
  - image processing
  - 3D computer graphics
  - physically-based rendering
  - visualization
  - and, traditional art

Graduate Courses & Faculty Members

- Visualization (6406)
  - focuses on graphical techniques for data visualization that assist in the extraction of meaning from datasets
- Advanced Computer Graphics (6604)
  - covers topics in computer graphics, including rendering, geometric modeling, and computer animation
- Digital Image Processing (6602)
  - covers topics in digital picture processing such as visual perception, digitization, compression and enhancement

Characterizing a social bookmarking and tagging network

Evangelos Milios

Dalhousie Univ., Faculty of Computer Science

Joint research with
Ralitsa Angelova[1]
Marek Lipczak[2],
Paweł Prałat[2]

[1] Max Planck Institut für Informatik, Saarbrücken, Germany

[2] Dalhousie University

DALHOUSIE UNIVERSITY
Inspiring Minds

Outline

- Social bookmarking and collaborative tagging
- Friendship, common entity and similarity graphs
- k-core analysis
- Bookmark and tag distributions
- Friendships and bookmark/tag similarities
- Density properties
- Discussion

DALHOUSIE UNIVERSITY
Inspiring Minds

Social bookmarking and collaborative tagging systems

- On a system like del.icio.us, users:
  - store their personal bookmarks
  - tag them with keywords
  - establish one-directional friendship links to other users

DALHOUSIE UNIVERSITY
Inspiring Minds

Social bookmarking and collaborative tagging systems

- On a system like del.icio.us, users:
  - store their personal bookmarks
  - tag them with keywords
  - establish one-directional friendship links to other users
- We study the relation between friendship and similarity of bookmarks and tags

DALHOUSIE UNIVERSITY
Inspiring Minds

Social bookmarking and collaborative tagging systems

- On a system like del.icio.us, users:
  - store their personal bookmarks
  - tag them with keywords
  - establish one-directional friendship links to other users
- We study the relation between friendship and similarity of bookmarks and tags
- Study is centered on relations between users
  - friendship graphs
  - graphs based on common bookmarks / tags
  - graphs based on similarity of bookmarks / tags

DALHOUSIE UNIVERSITY
Inspiring Minds

Social bookmarking and collaborative tagging systems

- On a system like del.icio.us, users:
  - store their personal bookmarks
  - tag them with keywords
  - establish one-directional friendship links to other users
- We study the relation between friendship and similarity of bookmarks and tags
- Study is centered on relations between users
  - friendship graphs
  - graphs based on common bookmarks / tags
  - graphs based on similarity of bookmarks / tags
- Questions:
  - Do friends share common interests?
  - Are tags user specific or generally meaningful?
  - What are the density properties of similarity graphs?

DALHOUSIE UNIVERSITY
Inspiring Minds

Motivation

- Clustering users in social bookmarking and collaborative tagging systems

DALHOUSIE UNIVERSITY
Inspiring Minds

Motivation

- Clustering users in social bookmarking and collaborative tagging systems
- Take into account
  - friendship links
  - bookmarks / tags in common
  - similar bookmark / tag sets

DALHOUSIE UNIVERSITY
Inspiring Minds

Motivation

- Clustering users in social bookmarking and collaborative tagging systems
- Take into account
  - friendship links
  - bookmarks / tags in common
  - similar bookmark / tag sets
- Need for modeling such networks
  - To further our understanding of their properties
  - To generate synthetic data sets for testing clustering algorithms

DALHOUSIE UNIVERSITY
Inspiring Minds

Previous Work

- Evolution Models of Flickr and Yahoo!360 (Kumar 2006)
- Search and ranking for social networks (Hotho 2006)
- Analysis of online social networks (Ahn 2007)
- Tagging distributions (small set of stable tags, long tail of idiosyncratic tags) (Golder 2005)
- Tag co-occurrence network detects tag spamming (Schmitz 2007)
- Social networks and the semantic web (Mika 2007)

DALHOUSIE UNIVERSITY
Inspiring Minds

Friendship graph

- Vertices are users

Friendship graph

- Vertices are users
- Edges are directed friendship links between users

DALHOUSIE UNIVERSITY
Inspiring Minds

Friendship graph

- Vertices are users
- Edges are directed friendship links between users
- We obtain undirected friendship graph by ignoring direction

DALHOUSIE UNIVERSITY
Inspiring Minds

Friendship graph

- Vertices are users
- Edges are directed friendship links between users
- We obtain undirected friendship graph by ignoring direction
- Bidirectional edges representing mutual friendship are included once

DALHOUSIE UNIVERSITY
Inspiring Minds

Common Entity Graphs

- Graphs implicitly defined by number of entities users have in common

Common Entity Graphs

- Graphs implicitly defined by number of entities users have in common
- Symmetric similarity metric
- Range from 0 to the maximum number of entities.

DALHOUSIE UNIVERSITY
Inspiring Minds

Common Entity Graphs

- Graphs implicitly defined by number of entities users have in common
- Symmetric similarity metric
- Range from 0 to the maximum number of entities.
- Only entities having over a certain user frequency (5) are considered

DALHOUSIE UNIVERSITY
Inspiring Minds

Common Entity Graphs

- Graphs implicitly defined by number of entities users have in common
- Symmetric similarity metric
- Range from 0 to the maximum number of entities.
- Only entities having over a certain user frequency (5) are considered
- Two types of graphs:
  - Common bookmark graph (entities are bookmarks)
  - Common tag graph (entities are tags)

DALHOUSIE UNIVERSITY
Inspiring Minds

Similarity graphs

- Vertices connected by undirected weighted edge that reflects the cosine similarity between the entity vectors of the corresponding users

DALHOUSIE UNIVERSITY
Inspiring Minds

Similarity graphs

- Vertices connected by undirected weighted edge that reflects the cosine similarity between the entity vectors of the corresponding users
- Vector space defined by the set of entities over the entire system
- Only entities having over a certain user frequency (5) are included in the vector space

DALHOUSIE UNIVERSITY
Inspiring Minds

Similarity graphs

- Vertices connected by undirected weighted edge that reflects the cosine similarity between the entity vectors of the corresponding users
- Vector space defined by the set of entities over the entire system
- Only entities having over a certain user frequency (5) are included in the vector space
- There is no equivalent of “stop words” in bookmarks or tags

DALHOUSIE UNIVERSITY
Inspiring Minds

Similarity graphs

- Vertices connected by undirected weighted edge that reflects the cosine similarity between the entity vectors of the corresponding users
- Vector space defined by the set of entities over the entire system
- Only entities having over a certain user frequency (5) are included in the vector space
- There is no equivalent of “stop words” in bookmarks or tags
- Two types
  - Bookmark similarity graph. Weights are binary

DALHOUSIE UNIVERSITY *Inspiring Minds*

Similarity graphs

- Vertices connected by undirected weighted edge that reflects the cosine similarity between the entity vectors of the corresponding users
- Vector space defined by the set of entities over the entire system
- Only entities having over a certain user frequency (5) are included in the vector space
- There is no equivalent of “stop words” in bookmarks or tags
- Two types
  - Bookmark similarity graph. Weights are binary
  - Tag similarity graph. Weights are tf-idf scores for the tags.

DALHOUSIE UNIVERSITY
Inspiring Minds

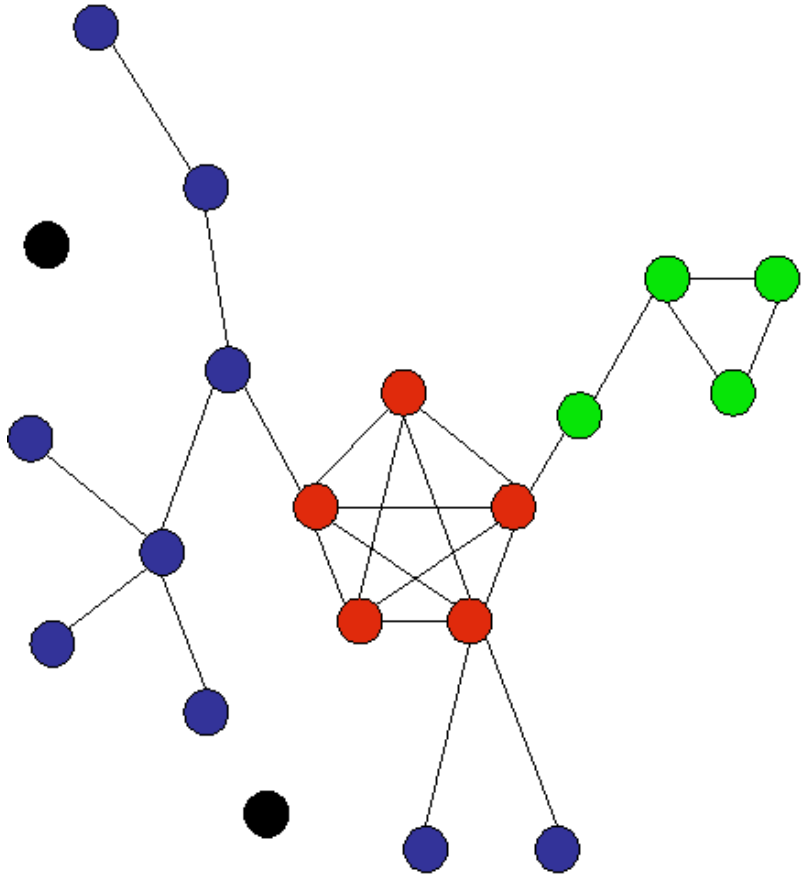
Similarity graphs

- Vertices connected by undirected weighted edge that reflects the cosine similarity between the entity vectors of the corresponding users
- Vector space defined by the set of entities over the entire system
- Only entities having over a certain user frequency (5) are included in the vector space
- There is no equivalent of “stop words” in bookmarks or tags
- Two types
  - Bookmark similarity graph. Weights are binary
  - Tag similarity graph. Weights are tf-idf scores for the tags.
- Both common entity graphs and similarity graphs are converted to **binary graphs** by removing all edges with weight below such a threshold that one million edges are kept.

DALHOUSIE UNIVERSITY
Inspiring Minds

k-cores

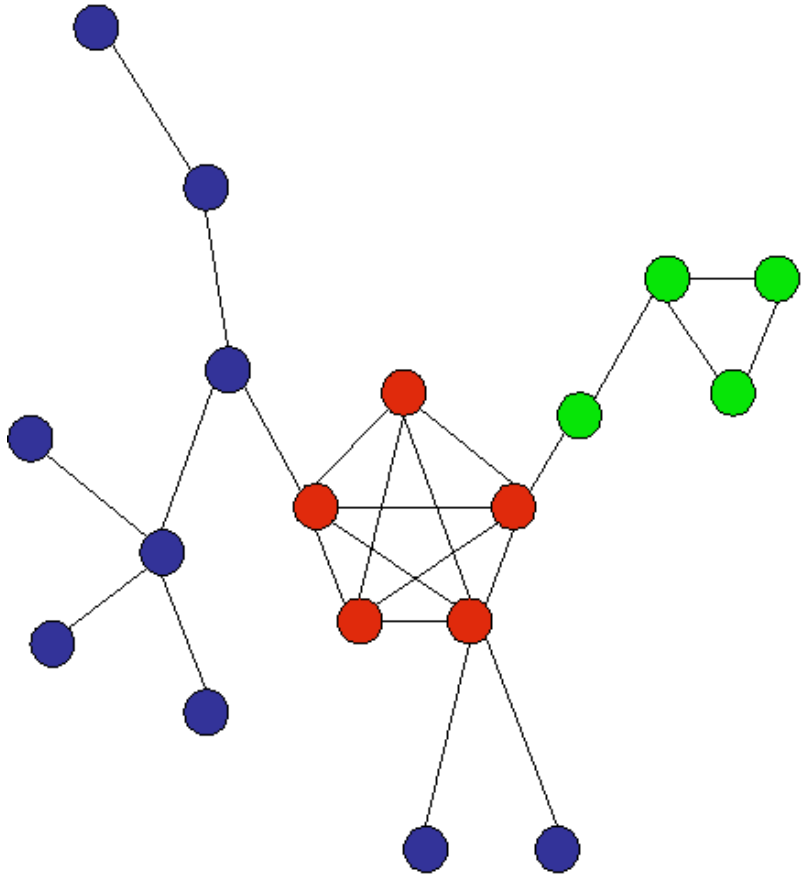
- Degree core of order k is the subgraph generated by recursively removing all nodes of degree less than k
- Here the $0 - core$ is the full graph
- It includes isolated vertices



DALHOUSIE UNIVERSITY
Inspiring Minds

k-cores

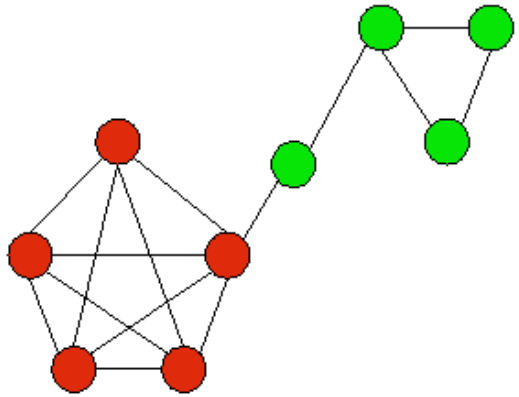
- Degree core of order k is the subgraph generated by recursively removing all nodes of degree less than k
- Here the $1 - core$ is the full graph with isolated vertices pruned



DALHOUSIE UNIVERSITY
Inspiring Minds

k-cores

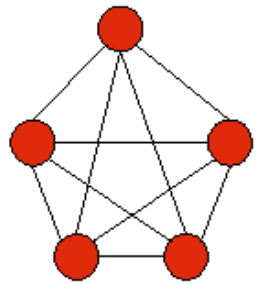
- Degree core of order k is the subgraph generated by recursively removing all nodes of degree less than k
- Here the $2-core$ prunes all tree structures



DALHOUSIE UNIVERSITY
Inspiring Minds

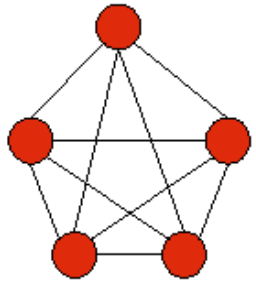
k-cores

- Degree core of order k is the subgraph generated by recursively removing all nodes of degree less than k
- Here is the $3-core$



k-cores

- Degree core of order k is the subgraph generated by recursively removing all nodes of degree less than k
- The $4-core$ is the same as the 3-core.



DALHOUSIE UNIVERSITY
Inspiring Minds

Density properties

- Clustering coefficient
  - for a vertex: number of actual edges between neighbours of a vertex as a fraction of the total potential number.
  - for a graph: average over all vertices (of degree > 1)

DALHOUSIE UNIVERSITY
Inspiring Minds

Density properties

- Clustering coefficient
  - for a vertex: number of actual edges between neighbours of a vertex as a fraction of the total potential number.
  - for a graph: average over all vertices (of degree > 1)
- K-core analysis
  - We produce a sequence of k-core graphs, with increasing k
  - We plot properties of these graphs as a function of k
    - ★ diameter of largest component
    - ★ size of largest component
    - ★ average distance between vertex pairs
    - ★ clustering coefficient
    - ★ number of components

DALHOUSIE UNIVERSITY Inspiring Minds

Density properties

- Scatter plots of the average clustering coefficient of the vertices of:
  - the same degree in the original graph versus degree
  - the largest component in the k-core sequence of graphs versus k

DALHOUSIE UNIVERSITY
Inspiring Minds

Exploration of our data set

- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed

Exploration of our data set

- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed
- $13,514$ users

DALHOUSIE UNIVERSITY
Inspiring Minds

Exploration of our data set

- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed
- $13,514$ users
- $4,574,587$ bookmarks

DALHOUSIE UNIVERSITY
Inspiring Minds

Exploration of our data set

- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed
- $13,514$ users
- $4,574,587$ bookmarks
- $47,807$ friendship connections

DALHOUSIE UNIVERSITY
Inspiring Minds

Exploration of our data set

- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed
- $13,514$ users
- $4,574,587$ bookmarks
- $47,807$ friendship connections
- $6,876$ of friendships are mutual

DALHOUSIE UNIVERSITY
Inspiring Minds

Exploration of our data set

- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed
- $13,514$ users
- $4,574,587$ bookmarks
- $47,807$ friendship connections
- $6,876$ of friendships are mutual
- $643,889$ tags in total

DALHOUSIE UNIVERSITY
Inspiring Minds

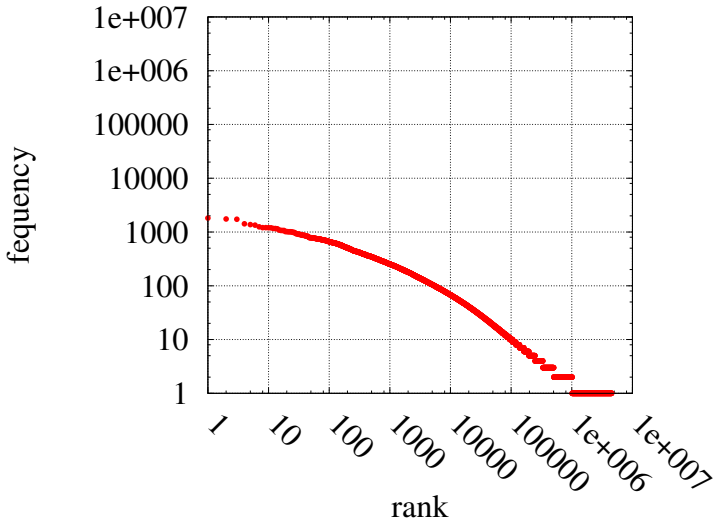
Exploration of our data set

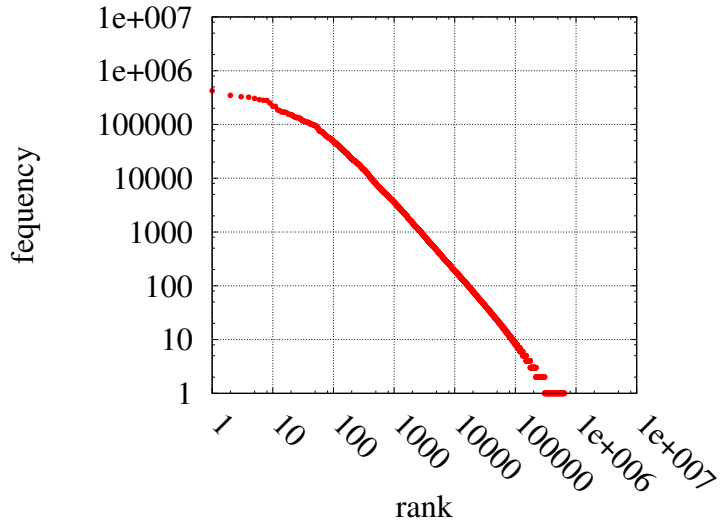
- Data set based on a Breadth first search of the friendship graph, starting with the user with the most friends as seed
- 13,514 users
- 4,574,587 bookmarks
- 47,807 friendship connections
- 6,876 of friendships are mutual
- 643,889 tags in total

Counts (in K)	urls	url domains	Tags	Wikipedia words
Total number of unique terms	4,575	1,106	644	4,098
Terms used more than once	1,017	483	303	1,978
Terms used only once	3,558	623	341	2,120

DALHOUSIE UNIVERSITY Inspiring Minds

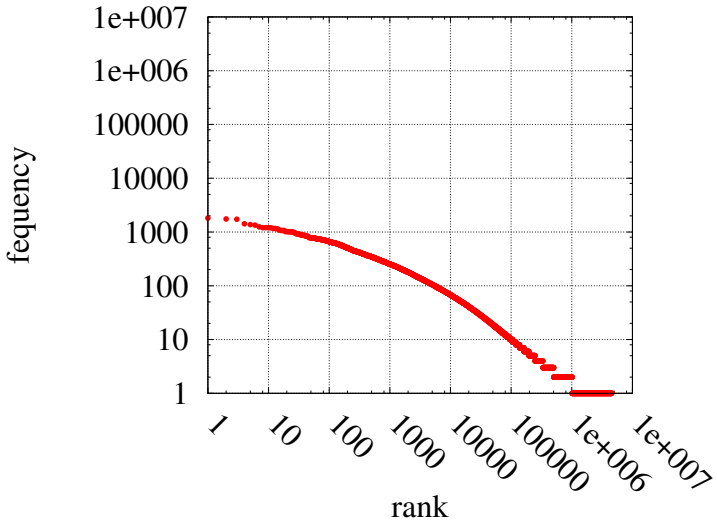
Zipf distributions for bookmarks and tags

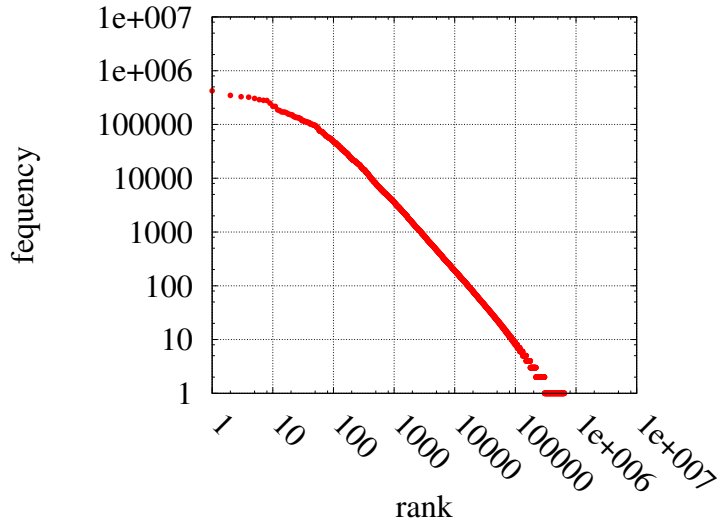




DALHOUSIE UNIVERSITY
Inspiring Minds

Zipf distributions for bookmarks and tags





- Tail is Zipf-like
- Head is flat (no “stop-word” behaviour)
- Large fraction appearing only once

DALHOUSIE UNIVERSITY
Inspiring Minds

Relating friendship with bookmark and tag similarity

pair average over →	Friends	Non-friends
k-common bookmarks	1.931	0.372
bookmark cosine sim	0.011	0.004
k-common tags	54.157	41.816
tag cosine sim	0.081	0.085

- Compare friend and non-friend pairs

Relating friendship with bookmark and tag similarity

pair average over →	Friends	Non-friends
k-common bookmarks	1.931	0.372
bookmark cosine sim	0.011	0.004
k-common tags	54.157	41.816
tag cosine sim	0.081	0.085

- Compare friend and non-friend pairs
- Friends have stronger connections over bookmarks

DALHOUSIE UNIVERSITY
Inspiring Minds

Relating friendship with bookmark and tag similarity

pair average over →	Friends	Non-friends
k-common bookmarks	1.931	0.372
bookmark cosine sim	0.011	0.004
k-common tags	54.157	41.816
tag cosine sim	0.081	0.085

- Compare friend and non-friend pairs
- Friends have stronger connections over bookmarks
- Friends have similar connections over tags as non-tags

DALHOUSIE UNIVERSITY Inspiring Minds

Relating friendship with bookmark and tag similarity

pair average over $\longrightarrow$	Friends	Non-friends
k-common bookmarks	1.931	0.372
bookmark cosine sim	0.011	0.004
k-common tags	54.157	41.816
tag cosine sim	0.081	0.085

- Compare friend and non-friend pairs
- Friends have stronger connections over bookmarks
- Friends have similar connections over tags as non-tags
- Conjecture: most tags are individualized, pertaining to the particular ways a person organizes their bookmarks

DALHOUSIE UNIVERSITY
Inspiring Minds

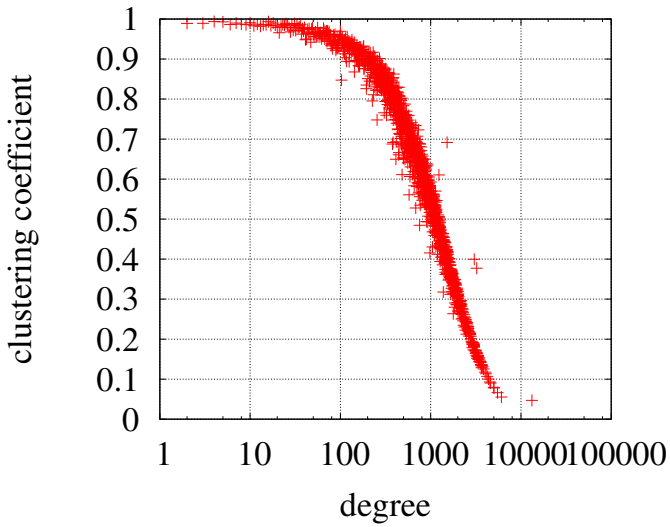
Relating friendship with bookmark and tag similarity

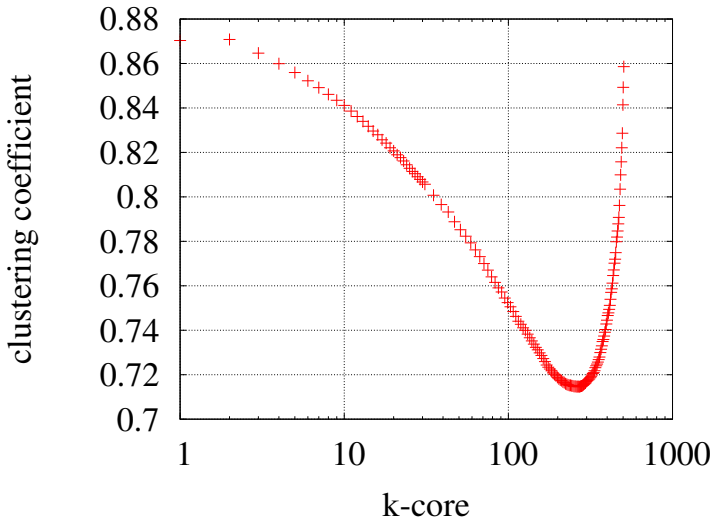
pair average over $\longrightarrow$	Friends	Non-friends
k-common bookmarks	1.931	0.372
bookmark cosine sim	0.011	0.004
k-common tags	54.157	41.816
tag cosine sim	0.081	0.085

- Compare friend and non-friend pairs
- Friends have stronger connections over bookmarks
- Friends have similar connections over tags as non-tags
- Conjecture: most tags are individualized, pertaining to the particular ways a person organizes their bookmarks

DALHOUSIE UNIVERSITY
Inspiring Minds

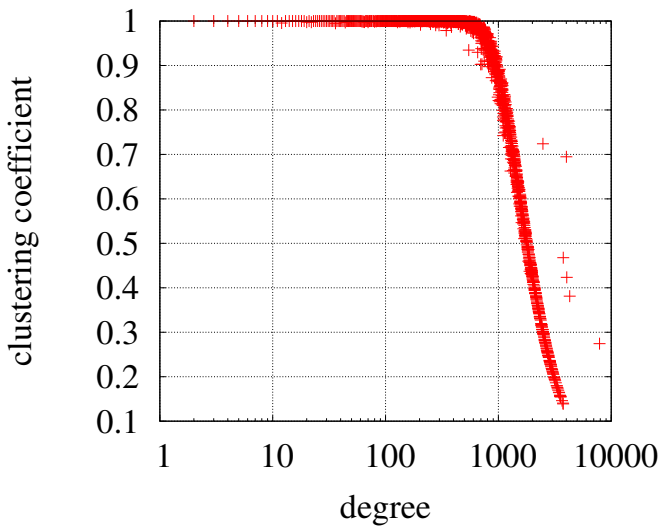
Density properties of common bookmark graphs

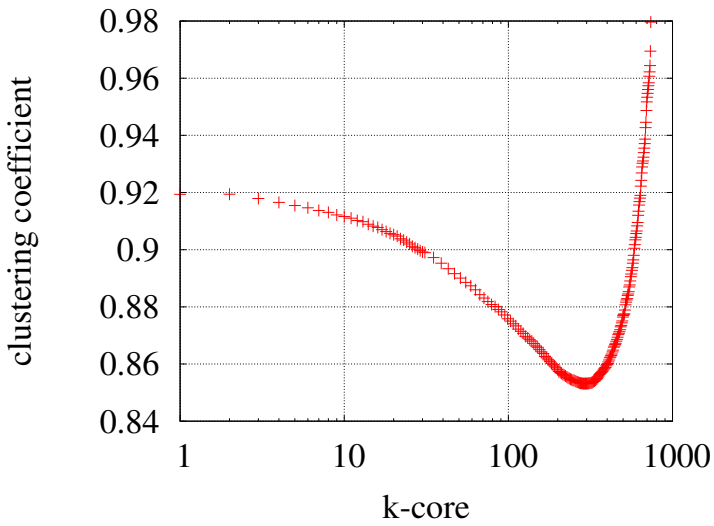




DALHOUSIE UNIVERSITY
Inspiring Minds

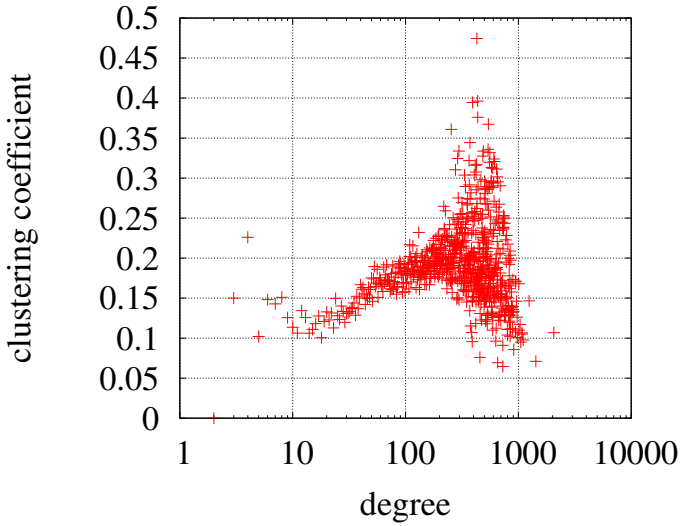
Density properties of common tag graphs

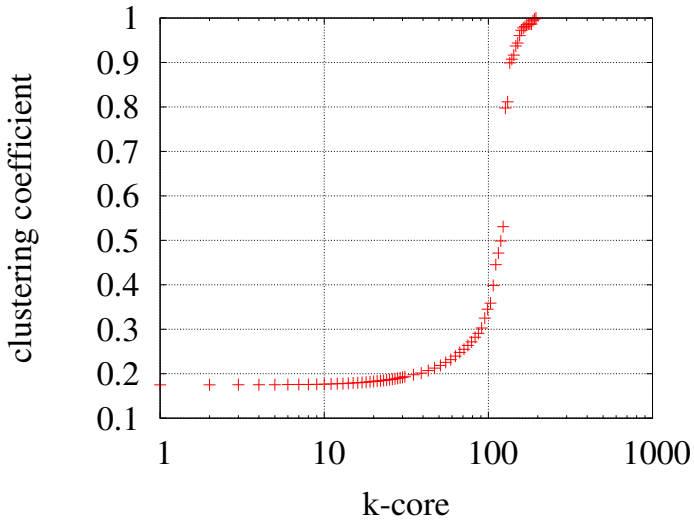




DALHOUSIE UNIVERSITY
Inspiring Minds

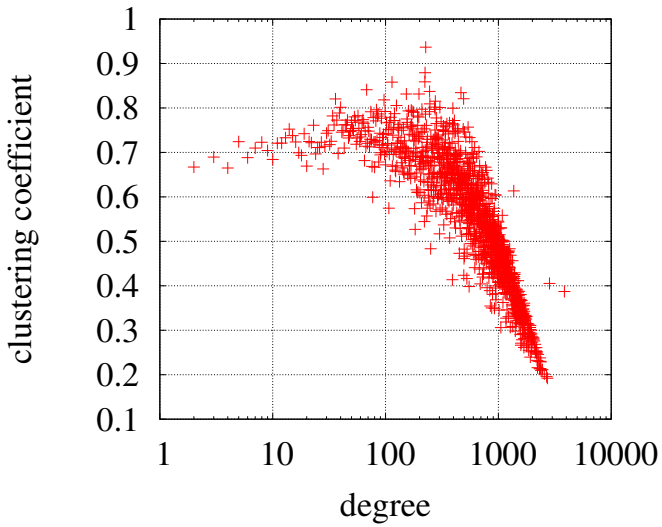
Density properties of bookmark similarity graphs

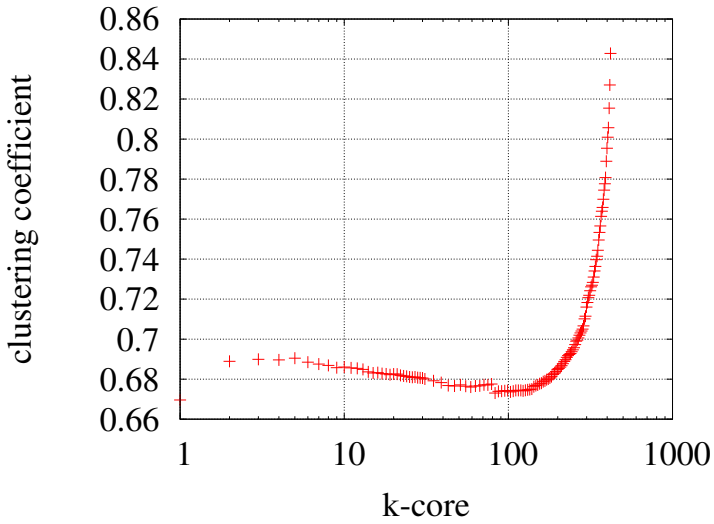




DALHOUSIE UNIVERSITY
Inspiring Minds

Density properties of tag similarity graphs





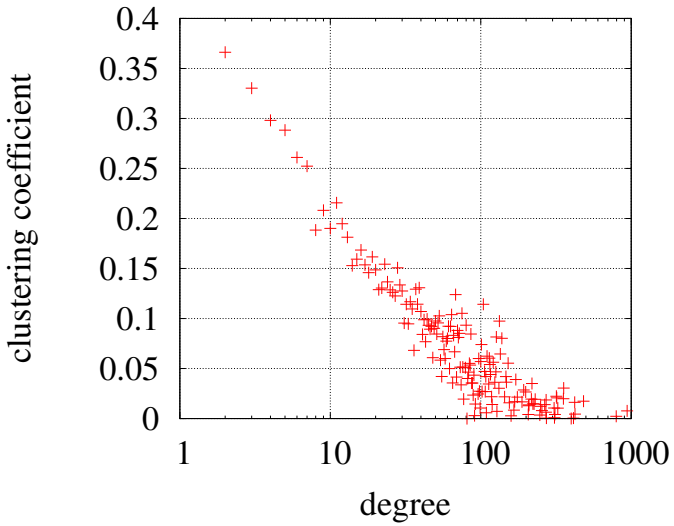
DALHOUSIE UNIVERSITY
Inspiring Minds

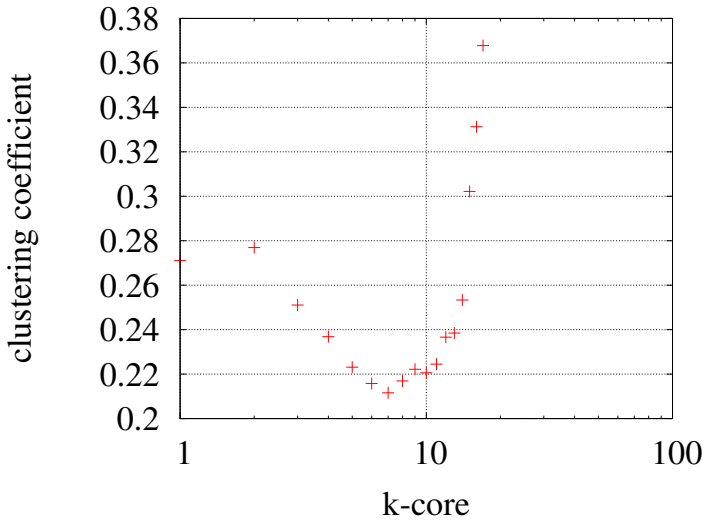
Observations

- Low-degree vertices are dropped first as k increases
- Such vertices have high clustering coefficients, hence
- average clustering coefficient drops
- As k keeps increasing, densification process prevails
- average clustering coefficient increases

DALHOUSIE UNIVERSITY
Inspiring Minds

Density properties of friendship graph





- Low degree vertices have high clustering coefficients, i.e.
- friends of users with few friends are friends themselves
- friends of users with large degrees are generally not connected

DALHOUSIE UNIVERSITY
Inspiring Minds

How do plots look from random graphs with power-law degree distribution?

- Constant!

How do plots look from random graphs with power-law degree distribution?

- Constant!
- The fact that two vertices are friends of a third vertex does not affect the probability of them being linked

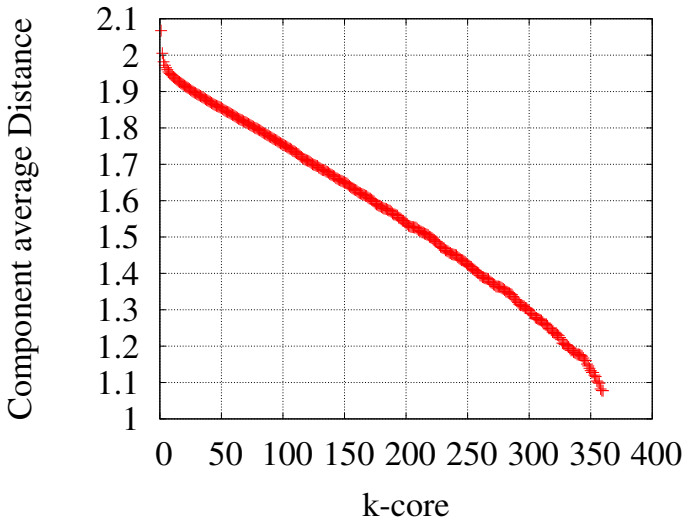
DALHOUSIE UNIVERSITY
Inspiring Minds

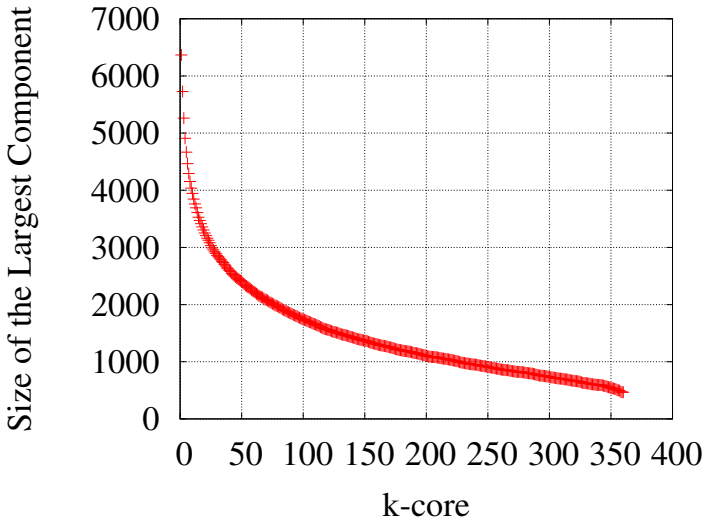
How do plots look from random graphs with power-law degree distribution?

- Constant!
- The fact that two vertices are friends of a third vertex does not affect the probability of them being linked
- U-shaped curve of clustering coefficient vs k not consistent with binomial random graphs or random graphs with a power law degree distribution

DALHOUSIE UNIVERSITY
Inspiring Minds

Largest component of common bookmark graph





- Average distance between pairs of vertices
- Size as a function of k
- Size is close to k, hence close to a clique

DALHOUSIE UNIVERSITY
Inspiring Minds

Summary

- Friendship correlates well with common / similar bookmarks but not for tags

DALHOUSIE UNIVERSITY
Inspiring Minds

Summary

- Friendship correlates well with common / similar bookmarks but not for tags
- Majority of tags are user-specific tags

DALHOUSIE UNIVERSITY
Inspiring Minds

Summary

- Friendship correlates well with common / similar bookmarks but not for tags
- Majority of tags are user-specific tags
- Tags behave like words in text more than bookmarks

DALHOUSIE UNIVERSITY
Inspiring Minds

Summary

- Friendship correlates well with common / similar bookmarks but not for tags
- Majority of tags are user-specific tags
- Tags behave like words in text more than bookmarks
- No equivalent of stop words for tags or bookmarks

DALHOUSIE UNIVERSITY
Inspiring Minds

Summary

- Friendship correlates well with common / similar bookmarks but not for tags
- Majority of tags are user-specific tags
- Tags behave like words in text more than bookmarks
- No equivalent of stop words for tags or bookmarks
- Graphs deviate from power-law behaviour

DALHOUSIE UNIVERSITY
Inspiring Minds

Computational Intelligence

Edited by ALI GHORBANI and EVANGELOS MILIOS

- **Impact Factor: 1.415**
- **ISI Journal Citation Reports® Ranking:**

2006: 29/85 (Computer Science, Artificial Intelligence)

- **Frequency: Bi-Monthly**

- **FOCAL AREAS**
  - Machine learning , incl.
    - symbolic multi-strategy and cognitive learning
  - Web intelligence and semantic web
  - Discovery science and knowledge mining
  - Agents and multi-agent systems
  - Modern knowledge-based systems
  - Key application areas of AI
    - games, software engineering, e-commerce

WILEY-BLACKWELL

www.blackwellpublishing.com/coin

DALHOUSIE UNIVERSITY
Inspiring Minds