Modelling and Mining Networked Information Spaces

Evangelos Milios

Dalhousie Univ., Faculty of Computer Science

July 20, 2008

World

Copyright © 1988-2004 Microsoft Corp. and/or its suppliers. All rights reserved. http://www.microsoft.com/autoroute/
© 1993-2003 NAVTEQ and its suppliers. All rights reserved. © Crown Copyright 2004. All rights reserved. License number 100025500.

Copyright © 1988-2004 Microsoft Corp. and/or its suppliers. All rights reserved. http://www.microsoft.com/autoroute/
© 1993-2003 NAVTEQ and its suppliers. All rights reserved. © Crown Copyright 2004. All rights reserved. License number 100025500.

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

Outline

- Social Networks
- Networked Information Spaces (e.g. Citation graphs, Web graph)
- Social resource sharing and tagging systems
- Search
- Community formation
- Dynamics / growth
- Knowledge mining
- Challenges

Online Social Networks

- Online communities originally supported by
  - email ('70s)
  - mailing lists ('80s)
  - newsgroups ('80s)
  - blogs (early '00s)
  - wikis (early '00s)

Online Social Networks

- Online communities originally supported by
  - email ('70s)
  - mailing lists ('80s)
  - newsgroups ('80s)
  - blogs (early '00s)
  - wikis (early '00s)
- Contemporary online social networking services (mid '00s)
  - purpose is linking people
    (*linkedin, facebook*)
  - purpose is to share resources,
    linking people is an extra
    (*flickr, del.ic.ious, yahoo!360, myspace*)

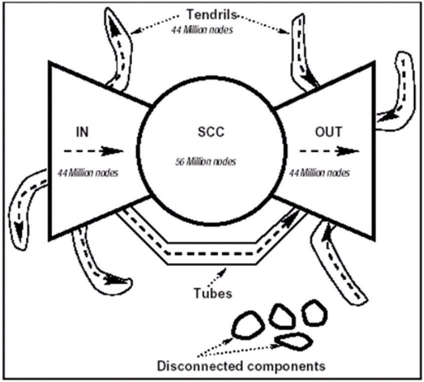
Networked Information Spaces

- Resources (Documents)
- Explicit links between resources
- *Organically grown* by a distributed community of contributors working independently
- Examples
  - Citation graph of research / patent literature
  - Gopher
  - World Wide Web
  - Common Law
  - Peer-to-peer information networks

Structure of Networked Information Spaces

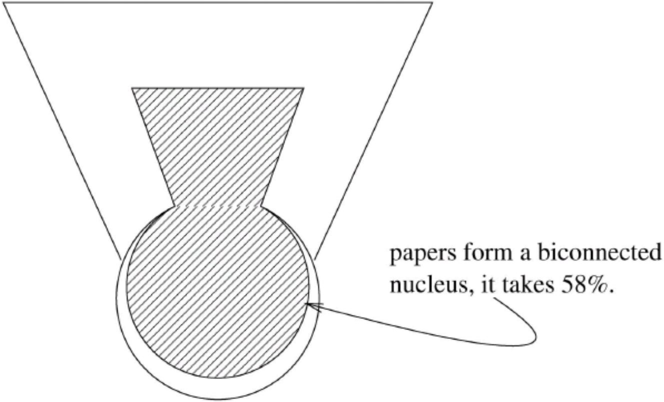
- Small world graphs
  - Short diameter
  - Small degree of separation (average distance between any two nodes)
- Power-law degree distributions (scale-free)
- “Strongly” connected (hard to break up by removing nodes)
- A tightly connected core plus small components

The bowtie model of the Web [1]



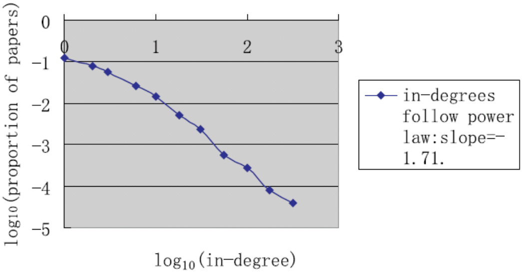
[1] Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, Wiener: Graph structure in the web, WWW-9, 1999

A model for the citation graph [2]

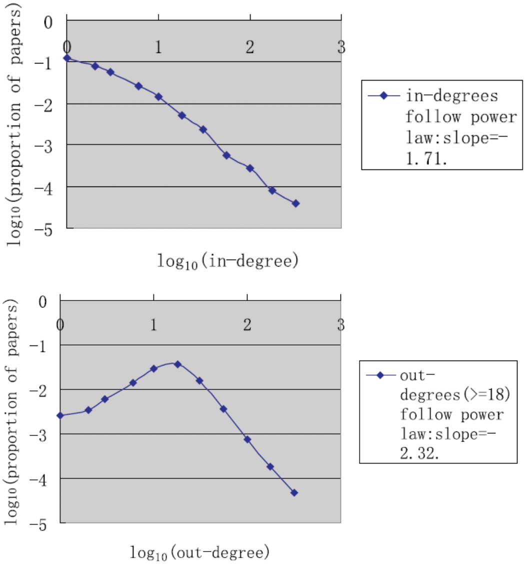


[2] An, Janssen, Milios: Characterizing and Mining the Citation Graph of Computer Science, Knowledge and Information Systems, 2004

Power laws in the citation graph



Power laws in the citation graph



Citation graph is hard to break up

Sizes of the largest Weakly Connected Components(WCCs) when nodes with in-degree at least k are removed from the giant connected component of union citation graph.

size of graph	50,228							
k	200	150	100	50	10	5	4	3
size of graph after removing	50,222	50,215	50,152	49,775	46,850	43,962	42,969	41,246
size of largest WCC	50,107	49,990	48,973	43,073	26,098	14,677	9,963	1,140

DALHOUSIE UNIVERSITY *Inspiring Minds*

Citation graph is hard to break up

Sizes of the largest Weakly Connected Components(WCCs) when nodes with in-degree at least k are removed from the giant connected component of union citation graph.

size of graph	50,228							
k	200	150	100	50	10	5	4	3
size of graph after removing	50,222	50,215	50,152	49,775	46,850	43,962	42,969	41,246
size of largest WCC	50,107	49,990	48,973	43,073	26,098	14,677	9,963	1,140

Sizes of the largest Weakly Connected Components(WCCs) when nodes with out-degree at least k are removed from the giant connected component of union citation graph.

size of graph	50,228							
k	200	150	100	50	10	5	4	3
size of graph after removing	50,225	50,225	50,224	50,205	48,061	43,964	42,238	39,622
size of largest WCC	50,202	50,202	50,198	50,131	46,092	37,556	33,279	26,489

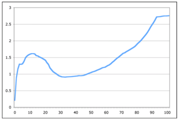
Dynamics of citation networks [3]

- Densification:
  - Average degree increases over time
  - according to a power law, $e(t) \propto n(t)^a$, $e(t), n(t)$ number of edges/nodes at time t
- Shrinking diameter as network grows
- Experimental data from
  - arXiv citation graph
  - patent citation graph
  - autonomous network graph
  - affiliation graphs (bipartite author/paper graphs)

[3] Leskovec, Kleinberg, Faloutsos: Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, KDD 2005

Dynamics of social networks [4]

- Density



- Structure
  - Giant component
  - Isolated star-shaped communities
  - Singletons
- Experimental data from Flickr, Yahoo 360!

4 Kumar, Novak, Tomkins: Structure and Evolution of Online Social Networks, KDD 2006

Modelling Evolution by Biased Preferential Attachment[5]

- Three types of users (user added at each timestep)
  - Passive (no activity)
  - Inviters (pull together an off-line community)
  - Linkers (full participants)
- Edges are added at each timestep
  - Source chosen at random from inviters/linkers with probability equal to degree (preferential attachment)
  - If source is inviter, a new node is created as destination
  - if source is a linker, an existing node is chosen by preferential attachment from inviters/linkers

[5] Kumar, Novak, Tomkins: Structure and Evolution of Online Social Networks, KDD 2006

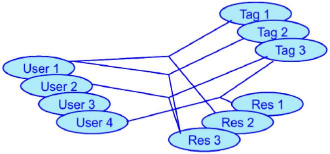
Social resource sharing and tagging systems

- On a social bookmarking and tagging system, users:
  - store resources (bookmarks, photos, music, video, publications, etc.)
  - tag them with keywords
  - establish one-directional friendship/contact links to other users

[6] Hotho, Jäschke, Schmitz, Stumme: Trend Detection in Folksonomies, SAMT 2006

Social resource sharing and tagging systems

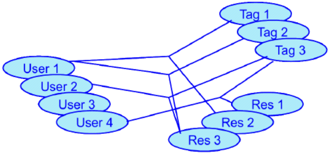
- On a social bookmarking and tagging system, users:
  - store resources (bookmarks, photos, music, video, publications, etc.)
  - tag them with keywords
  - establish one-directional friendship/contact links to other users
- Tripartite (hyper)graph structure:



[6] Hotho, Jäschke, Schmitz, Stumme: Trend Detection in Folksonomies, SAMT 2006

Social resource sharing and tagging systems

- On a social bookmarking and tagging system, users:
  - store resources (bookmarks, photos, music, video, publications, etc.)
  - tag them with keywords
  - establish one-directional friendship/contact links to other users
- Tripartite (hyper)graph structure:



- **Tag assignment**: (u, t, r), where u: user, t: tag, r: resource [6]

[6] Hotho, Jäschke, Schmitz, Stumme: Trend Detection in Folksonomies, SAMT 2006

Modelling and Mining social resource sharing and tagging systems

- Navigation
- Search and ranking
- Hypergraph structure
- Relation between friendship/contact links and resource/tag similarities
- Communities of users
- Taxonomies of tag concepts / topics (folksonomies)
- Trend detection, evolution, dynamics

Ranking: The FolkRank algorithm [7]

- Adapted PageRank:
  - Transform tripartite hypergraph into an undirected, weighted tripartite graph
  - Mutual reinforcement: Important resources tagged with important tags by important users

[7] Hotho, Jäschke, Schmitz, Stumme: FolkRank: A Ranking Algorithm for Folksonomies, FGIR 2006

Ranking: The FolkRank algorithm [7]

- Adapted PageRank:
  - Transform tripartite hypergraph into an undirected, weighted tripartite graph
  - Mutual reinforcement: Important resources tagged with important tags by important users
  - Random surfer model for setting weight vector $\vec{w}$ as a fixed point of iteration $\vec{w}$: $\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p}$, where A is the row-normalized adjacency matrix of graph, $\vec{p}$ is a personalization or topic-specific bias, and $\|\vec{w}\| = \|\vec{p}\|$.
  - For no bias, $\vec{p} = (1, 1, ..., 1)^T$

[7] Hotho, Jäschke, Schmitz, Stumme: FolkRank: A Ranking Algorithm for Folksonomies, FGIR 2006

Ranking: The FolkRank algorithm [7]

- Adapted PageRank:
  - Transform tripartite hypergraph into an undirected, weighted tripartite graph
  - Mutual reinforcement: Important resources tagged with important tags by important users
  - Random surfer model for setting weight vector $\vec{w}$ as a fixed point of iteration $\vec{w}$: $\vec{w} \leftarrow dA\vec{w} + (1 - d)\vec{p}$, where A is the row-normalized adjacency matrix of graph, $\vec{p}$ is a personalization or topic-specific bias, and $\|\vec{w}\| = \|\vec{p}\|$.
  - For no bias, $\vec{p} = (1, 1, ..., 1)^T$
- FolkRank
  - Compute $\vec{w}_0$ as fixed point with $d = 1$
  - Compute $\vec{w}_1$ as fixed point with $d < 1$
  - Final weight vector is $\vec{w} := \vec{w}_1 - \vec{w}_0$

DALHOUSIE UNIVERSITY Inspiring Minds

[7] Hotho, Jäschke, Schmitz, Stumme: FolkRank: A Ranking Algorithm for Folksonomies, FGIR 2006

Ranking: The FolkRank algorithm [7]

- Adapted PageRank:
  - Transform tripartite hypergraph into an undirected, weighted tripartite graph
  - Mutual reinforcement: Important resources tagged with important tags by important users
  - Random surfer model for setting weight vector $\vec{w}$ as a fixed point of iteration $\vec{w}$: $\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p}$, where A is the row-normalized adjacency matrix of graph, $\vec{p}$ is a personalization or topic-specific bias, and $\|\vec{w}\| = \|\vec{p}\|$.
  - For no bias, $\vec{p} = (1, 1, ..., 1)^T$
- FolkRank
  - Compute $\vec{w}_0$ as fixed point with $d = 1$
  - Compute $\vec{w}_1$ as fixed point with $d < 1$
  - Final weight vector is $\vec{w} := \vec{w}_1 - \vec{w}_0$
- Additional Uses of (Adapted) PageRank and FolkRank
  - Trend detection (how weights change for specific topics)
  - Community detection (influential users for specific topics)
  - Summarization

[7] Hotho, Jäschke, Schmitz, Stumme: FolkRank: A Ranking Algorithm for Folksonomies, FGIR 2006

DALHOUSIE UNIVERSITY
Inspiring Minds

Tag co-occurrence graphs [8]

- Vertices are tags
- Two tags are linked by an edge if a user has used them both on a resource

[8] Cattuto, Schmitz, Baldassarri, Servedio, Loreto, Hotho, Grahl, Stumme: Network Properties of Folksonomies, AI Comm. 2007

Tag co-occurrence graphs [8]

- Vertices are tags
- Two tags are linked by an edge if a user has used them both on a resource
- Strength of an edge, $w(i,j)$: number of tag assignments the two tags appear together
- Strength of a vertex s_i: sum of the strength of its incident edges

[8] Cattuto, Schmitz, Baldassarri, Servedio, Loreto, Hotho, Grahl, Stumme: Network Properties of Folksonomies, AI Comm. 2007

Tag co-occurrence graphs [8]

- Vertices are tags
- Two tags are linked by an edge if a user has used them both on a resource
- Strength of an edge, $w(i,j)$: number of tag assignments the two tags appear together
- Strength of a vertex s_i: sum of the strength of its incident edges
- Average nearest neighbour strength $S_{nn}(i)$: sum of the strengths of neighbour vertices of vertex i

[8] Cattuto, Schmitz, Baldassarri, Servedio, Loreto, Hotho, Grahl, Stumme: Network Properties of Folksonomies, AI Comm. 2007

Tag co-occurrence graphs [8]

- Vertices are tags
- Two tags are linked by an edge if a user has used them both on a resource
- Strength of an edge, $w(i, j)$: number of tag assignments the two tags appear together
- Strength of a vertex s_i: sum of the strength of its incident edges
- Average nearest neighbour strength $S_{nn}(i)$: sum of the strengths of neighbour vertices of vertex i
- Statistics of interest
  - Cumulative probability distribution of vertex strength
  - Scatter plot of s_i versus $S_{nn}(i)$

[8] Cattuto, Schmitz, Baldassarri, Servedio, Loreto, Hotho, Grahl, Stumme: Network Properties of Folksonomies, AI Comm. 2007

Tag spam detection

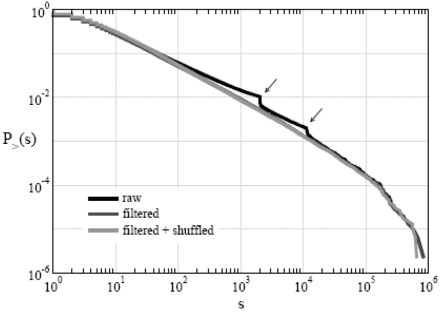


Fig. Cumulative strength distribution for the network of tag co-occurrence in del.icio.us. $P_>(s)$ is the probability of having a node with strength in excess of s. The black curve corresponds to the whole co-occurrence network. The two steps indicated by arrows correspond to an excess of links with a specific weight and can be related to spamming activity. Excluding from the analysis all posts with more than 50 tags removes the steps (dark gray). Shuffling the tags contained in posts (light gray) does not affect significantly the cumulated weight distribution. This proves that such a distribution is uniquely determined by tag frequencies within the folksonomy, and not by the semantics of co-occurrence.

Spikes reveal spamming behaviour

Relation of a vertex strength to that of its neighbours

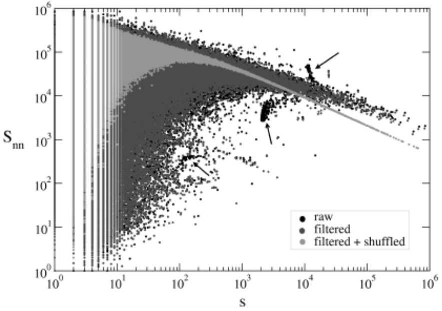


Fig. Average nearest-neighbor strength S_{nn} of nodes (tags) in relation to the node (tag) strengths s, in del.icio.us. Black dots correspond to the whole co-occurrence network. Assortative behavior is observed for low values of the strength s, while disassortative behavior is visible for high values of s. A few clusters (indicated by arrows) stand out from the main cloud of data points. As in Fig. 12, such anomalies correspond to spamming activity and can be removed by filtering out posts containing an excessive number of tags (dark grey). Shuffling the tags (light grey) affects dramatically the distribution of data points: this happens because the average nearest-neighbor strength of nodes is able to probe the local structure of the network of co-occurrence beyond the pure frequency effects, and is sensitive to patterns of co-occurrence induced by semantics.

- Positive correlation for small strengths (assortative)
- Negative correlation for large strengths (disassortative)
- Spamming behaviour stands out from the main trend

User-centred properties of the `del.icio.us` network

- We study the relation between friendship and similarity of bookmarks and tags

User-centred properties of the `del.icio.us` network

- 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

User-centred properties of the `del.icio.us` network

- 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?

More discussion later today.

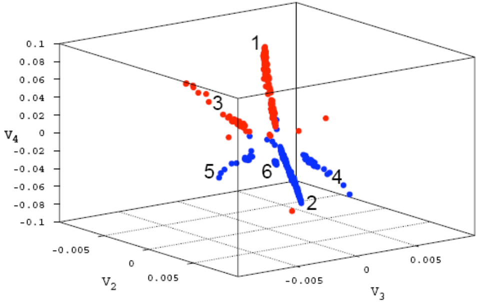
Resource-centred community formation in `del.icio.us` [9]

- Each resource is characterized by a tag cloud from the community of users
- Two resources are similar if their tag clouds overlap (TF-IDF weights)
- Form similarity matrix W
- Raise similarities to a small power $\gamma = 0.1$ to reduce dynamic range
- Rearrange rows and columns to visually identify community structure
  - Form matrix $\hat{W}_{i,j} = (1 - \delta_{i,j}) W_{i,j}$
  - Form matrix $S_{i,j} = \delta_{i,j} \sum_j \hat{W}_{i,j}$
  - Form matrix $Q = S - \hat{W}$
  - Lowest non-zero eigenvalues of Q reveal community structure

DALHOUSIE UNIVERSITY Inspiring Minds

[9] Cattuto, Baldassarri, Servedio, Loreto: Emergent Community Structure in Social Tagging Systems, Adv.Phys. 2007

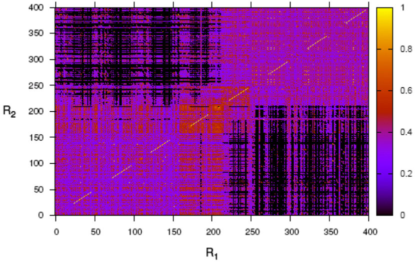
Results of community formation in `del.icio.us` [10]



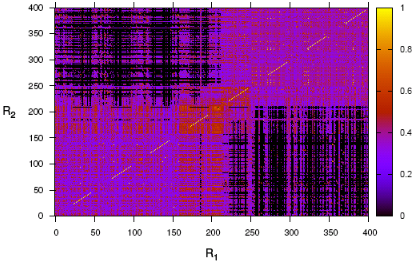
- Component values of the first three non-trivial eigenvectors
- Each point corresponds to an eigenvector component
- Coordinates are component values
- Clusters are visible

[10] Cattuto, Baldassarri, Servedio, Loreto: Emergent Community Structure in Social Tagging Systems, Adv.Phys. 2007

Reordered similarity matrix and tag clouds



Reordered similarity matrix and tag clouds



1 activism art blog bum bush creativity culture dvd economics flash freeware fun funny government history humor maps media money politics reference research software speechwriter statistics system:unfiled tools usa war windows

2 37signals art blog books css design development font fonts free graphics howto illustration inspiration photo photography photoshop portfolio productivity programming reference software system:unfiled themes tutorial tutorials typography web webdesign wordpress

3 activism blog blogs bush colbert comedy conservative culture election fraud freedom funny government grillo humor internet law libertarian maps media news political politics progressive science security system:unfiled usa video voting

4 art business color css design development flash free fun game games google graphics html inspiration patterns photography photos pricing reference resources search software stock system:unfiled tools web web2.0 webdesign webdev

5 ajax art awards blog blogger blogs color cool css design flash gallery graphics html images inspiration internet javascript lightbox politics portal portfolio reference system:unfiled templates tools web web2.0 webdesign webdev

6 ajax art blog books color css design desktop desktops development extension extensions firefox flash graphics icons illustration inspiration programming reference software system:unfiled technology tools typography wallpaper wallpapers web webdesign webdev

Challenges in community detection

- Clustering users

Challenges in community detection

- Clustering users
- Take into account
  - friendship links
  - bookmarks / tags in common
  - similar bookmark / tag sets

DALHOUSIE UNIVERSITY
Inspiring Minds

Challenges in community detection

- Clustering users
- 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

Challenges in Search [11]

- Content-based challenges
  - Short lifespan of content
  - Locality of interest
  - Vulnerability to spam

[11] Cho, Tomkins: Social Media and Search, IEEE Computer, Nov/Dec 2007

Challenges in Search [11]

- Content-based challenges
  - Short lifespan of content
  - Locality of interest
  - Vulnerability to spam
- System challenges
  - Access control
  - Distributed content (P2P)

[11] Cho, Tomkins: Social Media and Search, IEEE Computer, Nov/Dec 2007

Challenges: capturing emergent semantics [13]

- Hierarchies of tags
- Lightweight Ontology learning

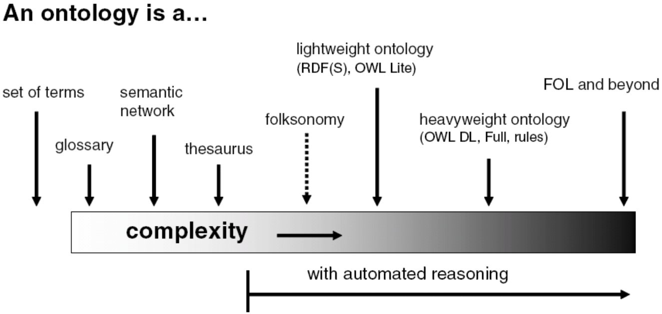
[12] Mika: Social Networks and the SemanticWeb, Springer, 2007, ch. 4

[13] Mika: Social Networks and the Semantic Web, Springer, 2007, ch. 9

Challenges: capturing emergent semantics [13]

- Hierarchies of tags
- Lightweight Ontology learning

What is a lightweight ontology? [12]



[12] Mika: Social Networks and the SemanticWeb, Springer, 2007, ch. 4

[13] Mika: Social Networks and the Semantic Web, Springer, 2007, ch. 9

Computational Intelligence

Edited by ALI GHORBANI and EVANGELOS MILIOS

- **Impact Factor: 1.972**
- **ISI Journal Citation Reports® Ranking:**

2007: 18/93 (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