

# Web Page Clustering Using Heuristic Search in the Web Graph

Ron Bekkerman

Shlomo Zilberstein

James Allan





# Example application

- Given a person name
- Find out *everything* about this person
  - What is available in the Web
- Possible solution:
  - Query a search engine with the person name
  - Retrieve documents
  - Cluster retrieved documents
    - Build *largest* clusters possible

# Query: "Michel Décary"

The screenshot shows a Windows Internet Explorer browser window with the title "Michel Décary" - Google Search - Windows Internet Explorer. The address bar contains the URL [http://www.google.com/search?as\\_q=&hl=en&num=10&btnG=Google+Search&as\\_epq=Michel+D%C3%A9cary&as\\_oq=&z](http://www.google.com/search?as_q=&hl=en&num=10&btnG=Google+Search&as_epq=Michel+D%C3%A9cary&as_oq=&z). The search bar contains the text "Michel Décary". The search results are displayed under the heading "Web" and show 10 results for "Michel Décary" (0.32 seconds).

The search results include:

- Stikeman Elliott LLP - Lawyer Profiles**: Michel Décary is a partner in the Montreal office of Stikeman Elliott, Fellow of the American College of Trial Lawyers, his practice focuses on litigation. [www.stikeman.com/en/avocats/profils/mtl/D\\_2006\\_216](http://www.stikeman.com/en/avocats/profils/mtl/D_2006_216) - Cached - Similar pages
- Stikeman Elliott S E N C R L s r l - Profils des avocats**: Michel Décary est associé au bureau de Montréal du cabinet Stikeman Elliott, Fellow de l'American College of Trial Lawyers, il a une pratique axée sur le... [www.stikeman.com/fr/avocats/profils/mtl/D\\_2006\\_216](http://www.stikeman.com/fr/avocats/profils/mtl/D_2006_216) - Cached - Similar pages
- Leiden Directory: Person**: Lectures at the ICC Corporate Governance College (Vrije University) and at the College des administrateurs de sociétés (Laval University). Michel Décary is... [www.leiden.nl/directory/person/prof/1147](http://www.leiden.nl/directory/person/prof/1147) - Not - Cached - Similar pages
- ZoomInfo Web Summary: Michel Décary**: With 20 years of natural language processing experience, Michel Décary co-founded Zoom Information and serves as chief scientist, responsible for developing... [www.zoominfo.com/MichelDecary](http://www.zoominfo.com/MichelDecary) - 28 - Cached - Similar pages
- ZoomInfo About Us: Management Team**: Michel Décary, Chief Scientist. Web summary: www.zoominfo.com/micheldecary With 20 years of natural language processing experience, Michel Décary co-founded... [www.zoominfo.com/AboutManagementTeam.aspx](http://www.zoominfo.com/AboutManagementTeam.aspx) - 49 - Cached - Similar pages
- DBLP: Michel Décary**: Michel Décary. List of publications from the DBLP Bibliography Server - PAC ... 1. Michel Décary. An Editor for the Explanatory and Combinatory Dictionary ... [www.informations.univie.ac.at/dblp/proceedings/2005/Decary/Michel.html](http://www.informations.univie.ac.at/dblp/proceedings/2005/Decary/Michel.html) - 26 - Cached - Similar pages
- RWCC - The Votes for Michel Décary**: Here are the votes we have recorded for Michel Décary. 1. The Myths & Legends of King Arthur & The Knights of the Round Table ... [www.rwcc.com/votes.asp?for\\_jeanmicheldecary](http://www.rwcc.com/votes.asp?for_jeanmicheldecary) - 28 - Cached - Similar pages
- An editor for the explanatory and combinatory dictionary of ...**: Michel Décary, Université de Montréal, Montréal, Québec, Canada, Guy Lapierre, Université de Montréal, Montréal, Québec, Canada ... [portal.slm.org/abstract/0170492079](http://portal.slm.org/abstract/0170492079) - Similar pages
- Web Page Clustering using Heuristic Search in the Web Graph**: The former PDF Adobe Acrobat - <http://www.slm.org> For instance, given a query Michel Décary, one can retrieve Web pages of at least three ... [www.slm.org/papers/papers/research.pdf](http://www.slm.org/papers/papers/research.pdf) - Similar pages
- Michel Décary - MP3**: Michel Décary remercie le Conseil des arts et des lettres du Québec de son appui financier. Répertoire des activités publiques. [pages.mvnet.net/decary/](http://pages.mvnet.net/decary/) - 49 - Cached - Similar pages

At the bottom of the search results, there is a "Go" button and a "Next" button. Below the search bar, there is a "Try Google Desktop" button with the text "search your computer as easily as you search the web."

The browser's status bar at the bottom shows "Internet" and "50%". The Windows taskbar at the very bottom shows the Start button, the search bar with "Michel Décary" - Goo..., and the system tray with "EN", "7:04 PM", and other icons.

# Query: "Michel Décary"

The collage consists of 12 browser window screenshots arranged in a 3x4 grid. The top row shows a Google search page with results for 'Michel Décary' and a 'Lexpert Directory' profile. The middle row shows a ZoomInfo web summary and a 'Mr. Michel Décary' profile page. The bottom row shows a music site profile and a 'Chansons en format MP3' page. The screenshots are connected by large curly braces on the right side, which are labeled 'Lawyer', 'Computer scientist', and 'Chanson singer'.

Lawyer

Computer scientist

Chanson singer



# Query: "Michel Décary"



## Education

McGill University, Master's Program in Law and Economics, course requirements completed (1969-1970), Université de Montréal (LL.L., 1967), and Université de Montréal (B.A., 1964).



Décary holds a B.A. in linguistics and an M.Sc. in computer science from Université de Montréal. He also pursued doctoral studies in computational linguistics at Université de Montréal and in applied linguistics at McGill University.



Mordu de mathématiques au niveau collégial, il découvre la linguistique formelle en 1977 et en fera sa nouvelle passion. Il obtient un baccalauréat spécialisé en linguistique de l'Université de Montréal en 1981. Après deux ans d'études vers l'obtention de la maîtrise, il change à nouveau de cap en découvrant le monde de l'informatique. Il refait l'équivalent du baccalauréat puis obtient une maîtrise en informatique (M.Sc., U de M, 1986). Co-fondateur et président de la firme informatique Machina Sapiens inc. (Le Correcteur 101) de 1985 à 1989, il poursuit en parallèle une carrière universitaire à l'Université McGill où il agit comme chercheur et directeur de projets informatiques dans le domaine du traitement des langues par ordinateur. Il enseigne pendant quatre ans l'enseignement des langues assisté par ordinateur et complète actuellement un doctorat dans ce domaine. Michel Décary est présentement chercheur pour le laboratoire d'informatique cognitive de la Télé-Université.



# Query: "Michel Décary"

Results 1 - 6 of about 7 linking to [www.cogilex.com](http://www.cogilex.com). (0.16 seconds)

**Founder**  
**Cogilex R&D inc. - Montreal, QC**

N'hésitez pas à m'écrire à [decary@cogilex.com](mailto:decary@cogilex.com)

Computer scientist

Chanson singer

The image shows a Google search interface with the query "Michel Décary". The search results are displayed in a grid format. The first result is a link to "www.cogilex.com", which is circled in red. Below the search results, there is a collage of various web pages related to Michel Décary, including a LinkedIn profile, a company page for Cogilex R&D inc., and a page with a photo of him. The text "Founder Cogilex R&D inc. - Montreal, QC" is overlaid on the collage. At the bottom of the collage, there is a text box that says "N'hésitez pas à m'écrire à decary@cogilex.com". To the right of the collage, there are two brackets with labels: "Computer scientist" and "Chanson singer".



# Query: "Michel Décary"

Lawyer

Computer scientist

+

Chanson singer



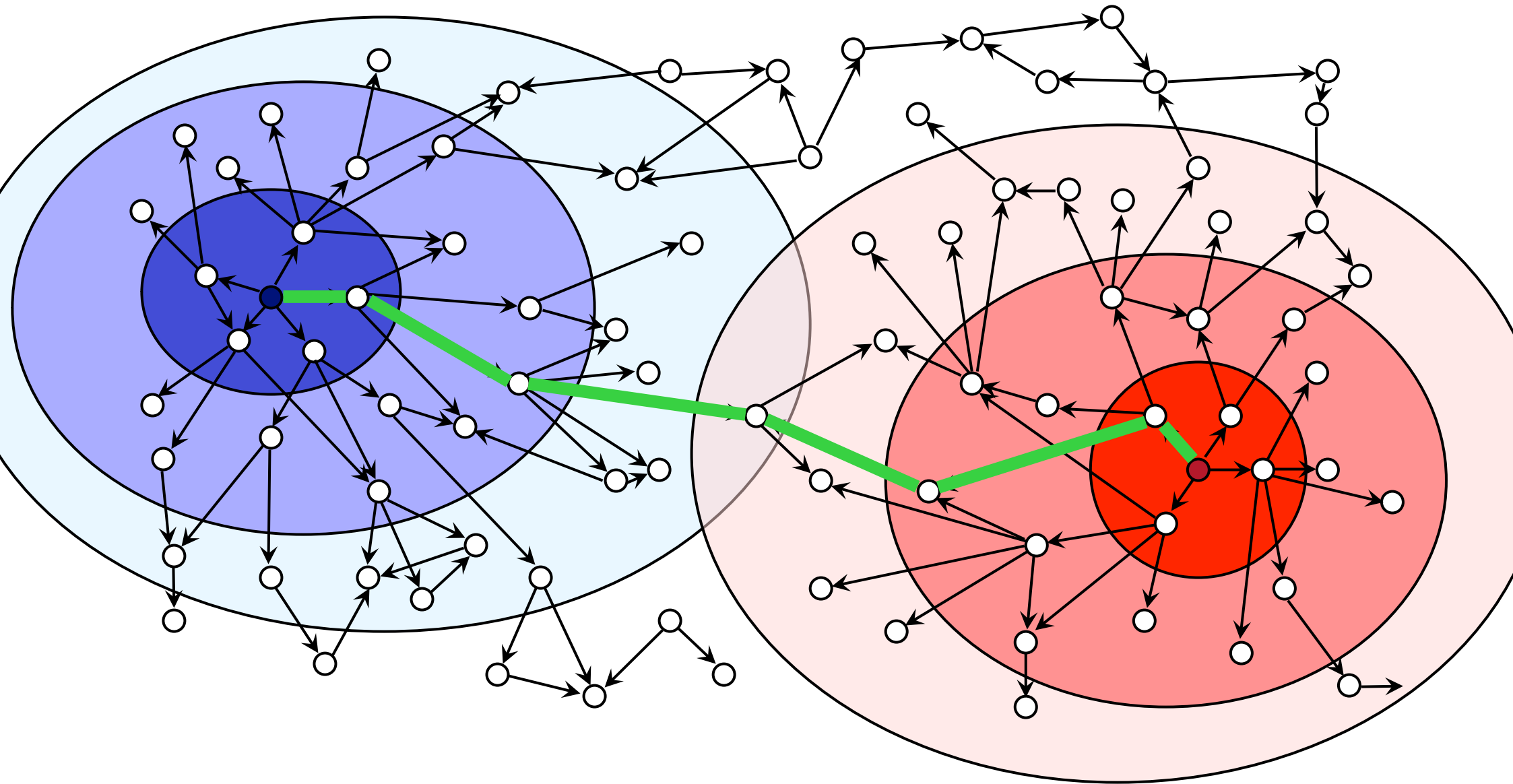
# Summary of observations

- *Topical* clustering is not enough
  - Although not to be ignored
- Web graph topology should be exploited
  - Close pages tend to be semantically related
    - There's a *reason* for hyperlinking page *A* → page *B*
    - Be careful: arbitrary connections exist as well
  - Apply *beam search* to find close pages
    - Use heuristics to prune undesired branches



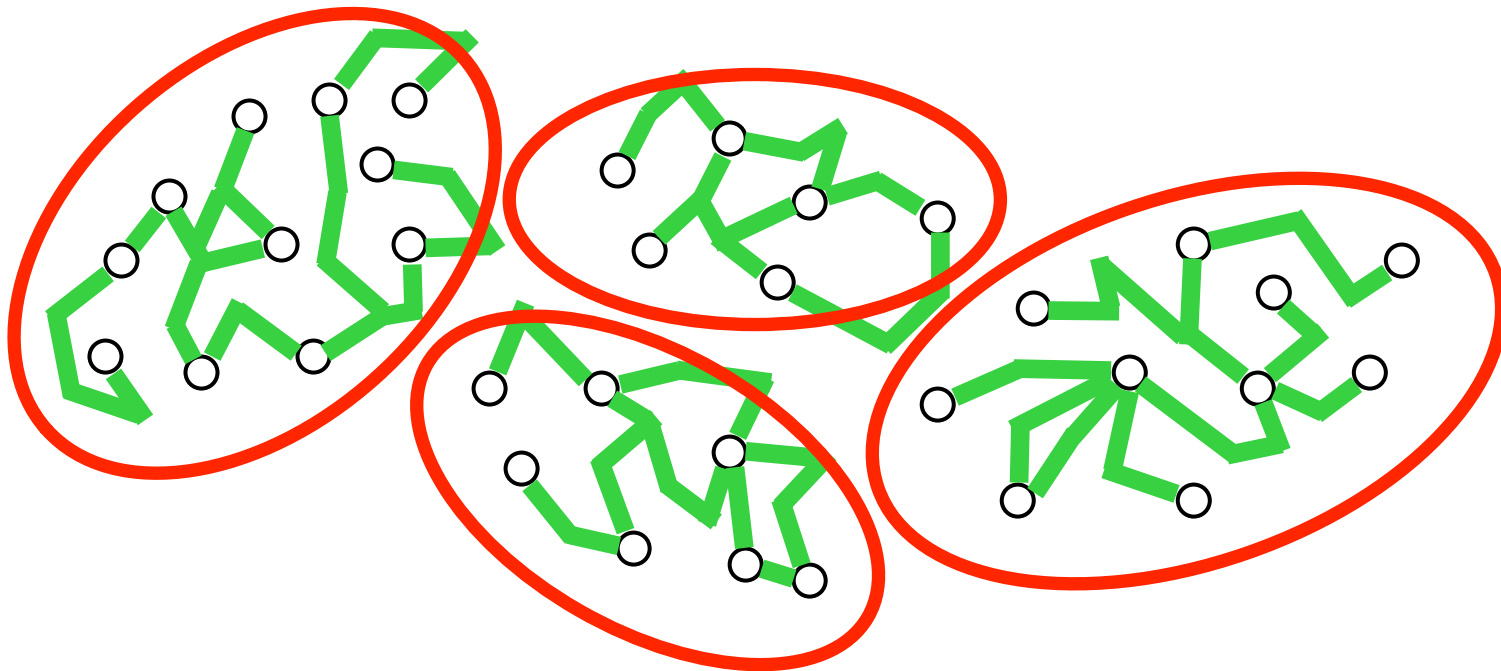


# Example: breadth first search



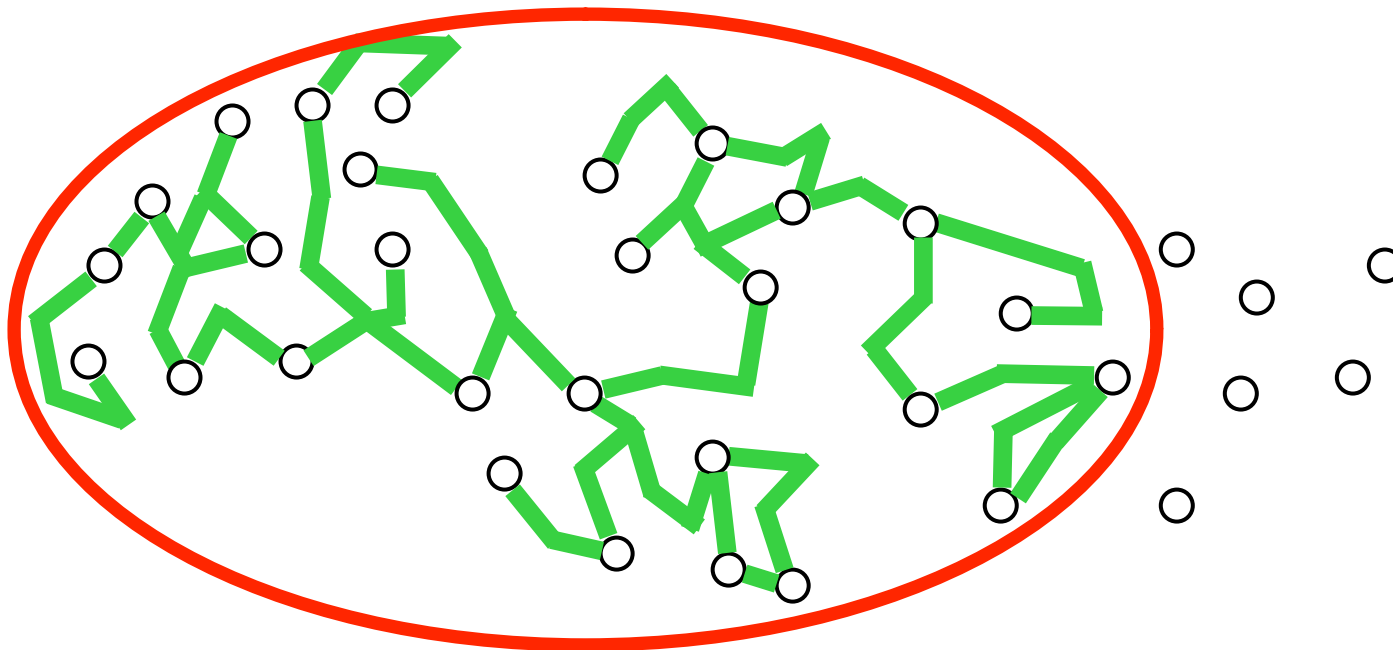
# Clustering by multi-agent search

- Each page is represented by a Web agent
  - Whose task is to traverse the Web graph
  - And meet as many other agents as possible



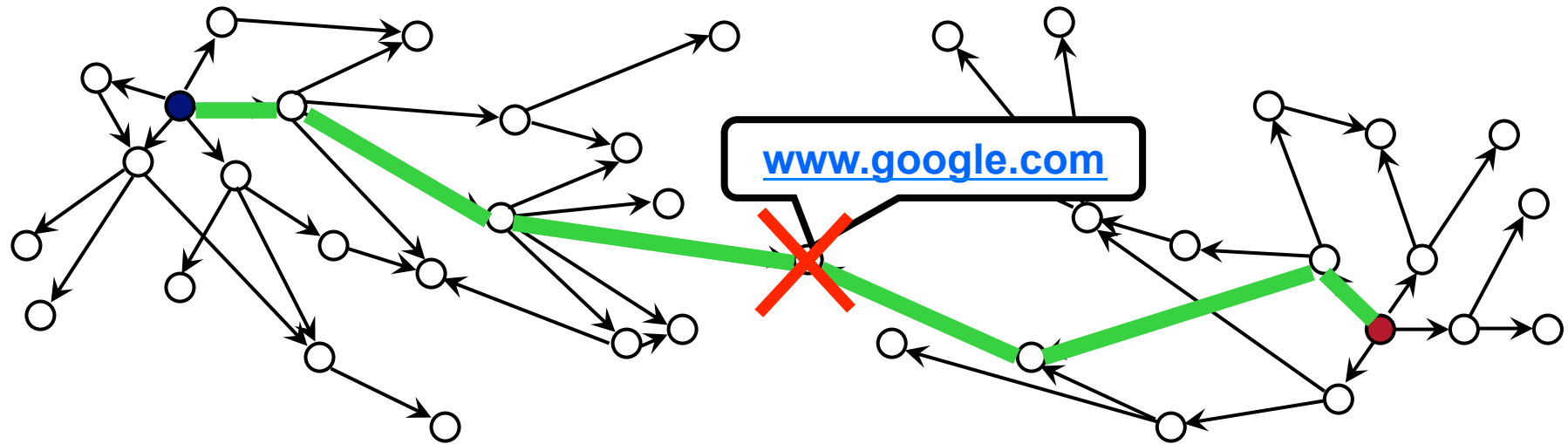
# Real-world case

- The Web is tightly interconnected
  - About 70% agents meet after 3 search iterations
  - Which is clearly an undesired outcome



# Heuristic 1

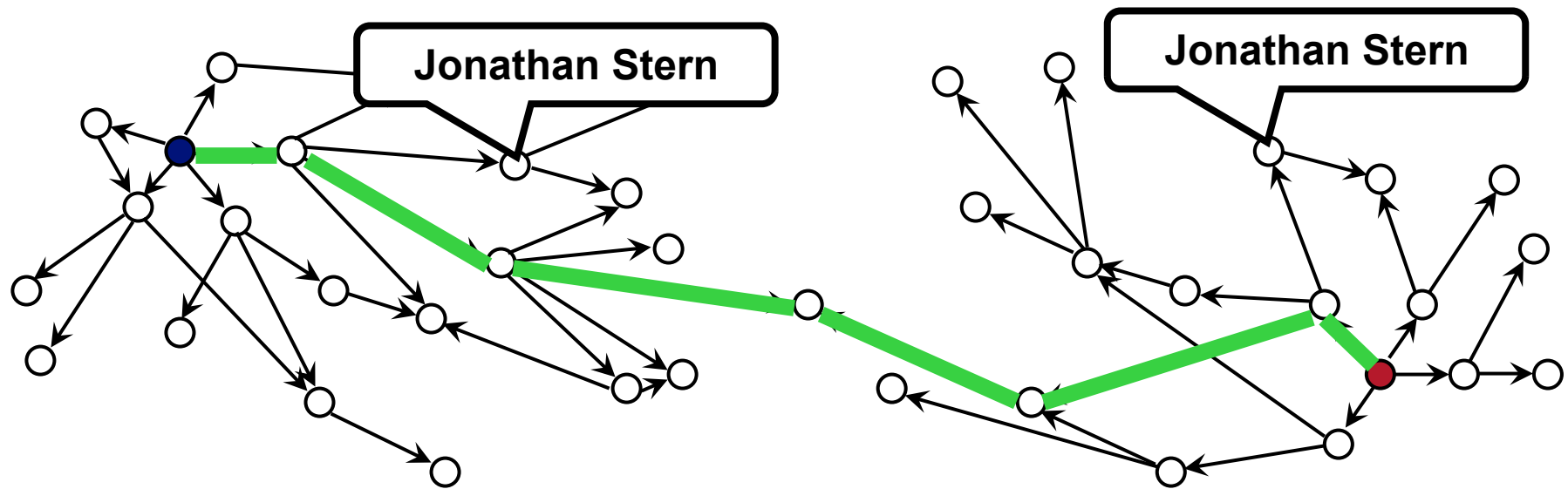
- Elimination of high-degree nodes
  - Both high in-degree and high out-degree
  - They often connect unrelated pages





# Heuristic 2

- Person name sharing
  - Expanded nodes share a hyperlink
  - AND a person name (ignore too popular names)



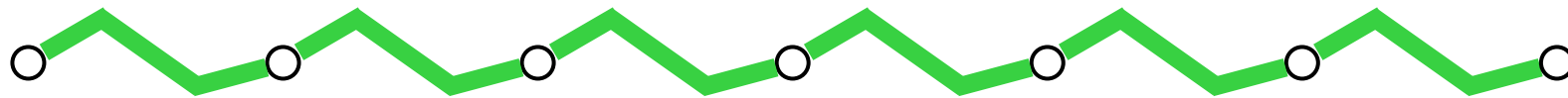


# Heuristic 3

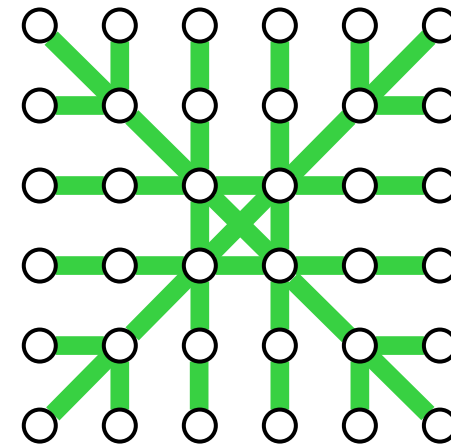
- Anchor text sharing
  - Anchor texts often summarize the content of hyperlinked pages
- Same idea as in person name sharing
  - But much simpler to implement
    - No sophisticated information extraction needed
    - Shallow parsing of HTML is enough
  - Again, ignore too frequent anchor texts
    - Like “*Contact Us*” or “*Copyright*”

# Algorithmic enhancement

- Unpleasant artifact: too long connections
  - Too weak semantic relationships



- Proposed solution: keep track of cluster's diameter
  - Start with a tightly connected component
  - Add pages found within one hop





# Experimentation domains

## ■ Web appearance disambiguation

*Bekkerman & McCallum, WWW-05*

- Given pages retrieved on  $N$  people names
  - From one social network
- Filter out pages that refer to their unrelated namesakes

## ■ Clustering of Web search results

*Hearst & Pedersen, SIGIR-96*

- Represent ranked lists of retrieved documents as clusters of semantically related documents



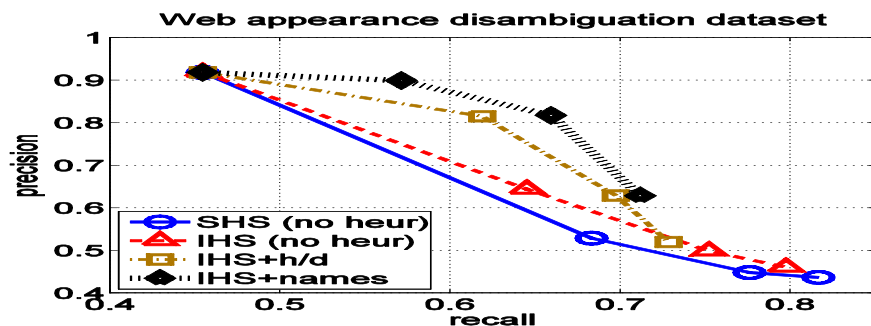


# Disambiguation dataset

- 12 names out of Melinda Gervasio's social network

<i>Personal name</i>	<i>Position</i>	<i>Pages</i>	<i>Namesakes</i>	<i>Relevant pages</i>
Adam Cheyer	SRI Manag	97	2	96
William Cohen	CMU Prof	88	10	6
Steve Hardt	SRI Eng	81	6	64
David Israel	SRI Manag	92	19	20
Leslie Pack Kaelbling	MIT Prof	89	2	88
Bill Mark	SRI Manag	94	8	11
Andrew McCallum	UMass Prof	94	16	54
Tom Mitchell	CMU Prof	92	37	15
David Mulford	Stanf Undergrad	94	13	1
Andrew Ng	Stanf Prof	87	29	32
Fernando Pereira	UPenn Prof	88	19	32
Lynn Voss	SRI Eng	89	26	1
	<b>OVERALL:</b>	<b>1085</b>	<b>187</b>	<b>420</b>

# Disambiguation results



- *h/d* – high degree node elimination
- *names* – person name heuristic
- Each point: one iteration of search
  - Only 2 iterations are enough
- *SHS* – sequential heuristic search (basic algorithm)
- *IHS* – incremental heuristic search
  - With the enhancement of diameter tracking

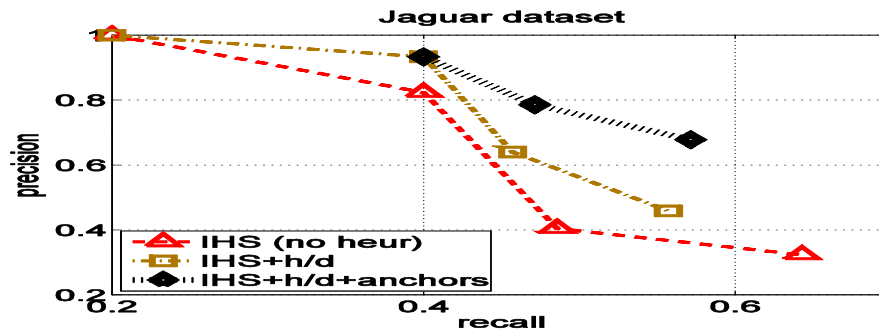


# Jaguar dataset

- 100 pages retrieved on query “*Jaguar*”
  - 23 different categories!
- The task is to build 3 clusters
  - Of cars, Mac OS and wild cats

<i>Category</i>	<i>Pages</i>	<i>Category</i>	<i>Pages</i>	<i>Category</i>	<i>Pages</i>
Car	36	Reef lodge	2	Atari game	5
MacOS	11	Book	1	Guitar	1
Wild cat	23	Singer	2	TV channel	1
Biotech firm	2	Emulator	2	Web designer	2
Youth org	1	Cornell project	2	E-commerce firm	1
Maya culture	1	Metal band	1	Game archive	1
Resin models	1	Movie	1	Aircraft	1
Web hosting	1	Photo gallery	1	<b>OVERALL:</b>	100

# Jaguar results



- SHS algorithm fails
  - 70 agents meet together
- *anchors* – anchor text heuristic
- Best performance:
  - High degree AND anchors heuristics



# Topical & topological clustering

- Build topical clusters [Bekkerman & McCallum, WWW-05](#)
- Enrich topical clusters with pages obtained by heuristic search based clustering

Web appearance disambiguation				Clustering of Web search results			
<i>Method</i>	<i>Precision</i>	<i>Recall</i>	<i>F-measure</i>	<i>Method</i>	<i>Precision</i>	<i>Recall</i>	<i>F-measure</i>
Topical	87.3%	71.3%	78.4%	Topical	75.0%	64.3%	69.2%
IHS (iteration 1)	89.9%	57.1%	69.9%	IHS (iteration 1)	93.3%	40.0%	56.0%
Hybrid (iteration 1)	84.5%	83.3%	<b>83.9%</b>	Hybrid (iteration 1)	77.1%	77.1%	<b>77.1%</b>
IHS (iteration 2)	81.7%	66.0%	73.0%	IHS (iteration 2)	78.6%	47.1%	58.9%
Hybrid (iteration 2)	78.5%	86.2%	82.2%	Hybrid (iteration 2)	72.7%	80.0%	76.2%

- Best performance: after one iteration of heuristic search only!



# Conclusion

- First application of heuristic search to the Web graph
  - Very simple algorithms / heuristics
  - Heuristic search theory yet to be applied
    - E.g., can an *admissible* heuristic be proposed?
- Search can be performed in real time!
  - Modern search engines store the link structure of most of the Web
  - Maximum 2 search iterations are enough
    - Fully distributable in a multi-agent fashion