

The Necessity of Mathematics

from

Google to Counterterrorism to Sudoku

Amy Langville

langvillea@cofc.edu

work supported by

NSF-CAREER-0546622,

NSA, DOEd, SAS, Semandex

Mathematics Department
College of Charleston
Charleston, SC

AMS Congressional Meeting

11/16/2006

The Message

- Mathematics is useful.
- Mathematical models don't care about scale or size of problem.
- Mathematical models are broadly applicable.
- Mathematical research is an inventive process.

Outline

- Sudoku
- Military Applications
 - planning flight paths
 - disabling and herding communication in networks
- Ranking Applications
 - ranking on the World Wide Web
- Clustering and Data Mining Applications
 - clustering the Enron email dataset
 - clustering on terrorist networks

Overriding Mathematical Techniques

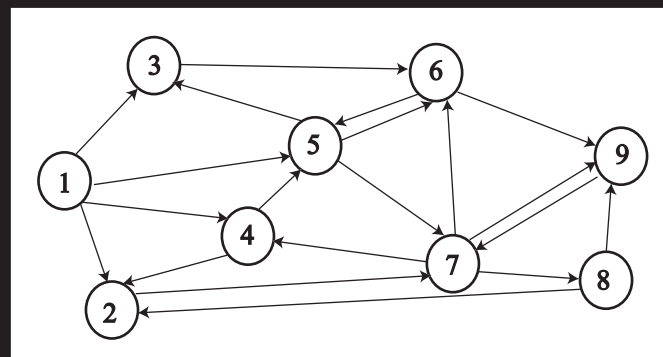
- Optimization

min/max Objective
subject to Constraint 1
Constraint 2
⋮

- Matrix Analysis

$$\begin{bmatrix} 4 & 5 & 0 & 1 \\ 2 & -1 & 3 & 1 \\ 0 & 7 & 0 & 1 \\ 2 & 0 & 0 & -3 \end{bmatrix}$$

- Graph Theory



Outline

- **Sudoku** optimization, matrices
- **Military Applications** optimization, graphs
 - planning flight paths
 - disabling and herding communication in networks
- **Ranking Applications** matrices, graphs
 - ranking on the World Wide Web
- **Clustering and Data Mining Applications** optimization, matrices, graphs
 - clustering the Enron email dataset
 - clustering on terrorist networks

Sudoku

Sudoku puzzle

		7	8		5	2		
8			6		4			5
	1			9			8	
4			2	8	9			7
5			7	6	1			2
	7			3			6	
3			1		6			4
		2	5		8	1		

Sudoku

Sudoku puzzle

		7	8		5	2		
8			6		4			5
	1			9			8	
4			2	8	9			7
5			7	6	1			2
	7			3			6	
3			1		6			4
		2	5		8	1		

Sudoku matrix

6	4	7	8	1	5	2	3	9
8	9	3	6	2	4	7	1	5
2	1	5	3	9	7	4	8	6
4	3	1	2	8	9	6	5	7
7	2	6	4	5	3	8	9	1
5	8	9	7	6	1	3	4	2
1	7	4	9	3	2	5	6	8
3	5	8	1	7	6	9	2	4
9	6	2	5	4	8	1	7	3

Sudoku

Sudoku puzzle

		7	8		5	2		
8			6		4			5
	1			9			8	
4			2	8	9			7
5			7	6	1			2
	7			3			6	
3			1		6			4
		2	5		8	1		

Sudoku matrix

6	4	7	8	1	5	2	3	9
8	9	3	6	2	4	7	1	5
2	1	5	3	9	7	4	8	6
4	3	1	2	8	9	6	5	7
7	2	6	4	5	3	8	9	1
5	8	9	7	6	1	3	4	2
1	7	4	9	3	2	5	6	8
3	5	8	1	7	6	9	2	4
9	6	2	5	4	8	1	7	3

Definition A $n \times n$ matrix is called a *Sudoku matrix* if:

1. n is a perfect square (e.g., 4, 9, 16, 25),
2. every row uses the integers 1 through n exactly once,
3. every column uses the integers 1 through n exactly once,
4. every submatrix uses the integers 1 through n exactly once.

Mathematical Model of Sudoku

Define:

$$x_{ijk} = \begin{cases} 1, & \text{if element } (i, j) \text{ of the } n \times n \text{ Sudoku matrix contains the integer } k \\ 0, & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \min \quad & \mathbf{0}^T \mathbf{x} \\ \text{s.t.} \quad & \sum_{i=1}^n x_{ijk} = 1, \quad j=1:n, k=1:n \quad (\text{only one } k \text{ in each column}) \end{aligned} \quad (1)$$

$$\sum_{j=1}^n x_{ijk} = 1, \quad i=1:n, k=1:n \quad (\text{only one } k \text{ in each row}) \quad (2)$$

$$\sum_{j=mq-m+1}^{mq} \sum_{i=mp-m+1}^{mp} x_{ijk} = 1, \quad k=1:n, p=1:m, q=1:m \quad (\text{only one } k \text{ in each submatrix}) \quad (3)$$

$$\sum_{k=1}^n x_{ijk} = 1 \quad i=1:n, j=1:n \quad (\text{every position in matrix must be filled}) \quad (4)$$

$$x_{ijk} = 1 \quad \forall (i, j, k) \in G \quad (\text{given elements } G \text{ in matrix are set "on"}) \quad (5)$$

$$x_{ijk} \in \{0, 1\} \quad (6)$$

Mathematical Model of Sudoku

Define:

$$x_{ijk} = \begin{cases} 1, & \text{if element } (i, j) \text{ of the } n \times n \text{ Sudoku matrix contains the integer } k \\ 0, & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \min \quad & \mathbf{0}^T \mathbf{x} \\ \text{s.t.} \quad & \sum_{i=1}^n x_{ijk} = 1, \quad j=1:n, k=1:n \quad (\text{only one } k \text{ in each column}) \end{aligned} \quad (1)$$

$$\sum_{j=1}^n x_{ijk} = 1, \quad i=1:n, k=1:n \quad (\text{only one } k \text{ in each row}) \quad (2)$$

$$\sum_{j=mq-m+1}^{mq} \sum_{i=mp-m+1}^{mp} x_{ijk} = 1, \quad k=1:n, p=1:m, q=1:m \quad (\text{only one } k \text{ in each submatrix}) \quad (3)$$

$$\sum_{k=1}^n x_{ijk} = 1 \quad i=1:n, j=1:n \quad (\text{every position in matrix must be filled}) \quad (4)$$

$$x_{ijk} = 1 \quad \forall (i, j, k) \in G \quad (\text{given elements } G \text{ in matrix are set "on"}) \quad (5)$$

$$x_{ijk} \in \{0, 1\} \quad (6)$$

Value of the Model

With a computer algorithm, we can solve any Sudoku puzzle, regardless of:

size n

number of givens

level of difficulty

9×9 puzzle takes 16.7 seconds to solve on desktop machine.

Unique Solution?

- Most puzzle creators do not check whether their puzzle has one unique solution.

Puzzle

2	2
3	4

Unique Solution?

- Most puzzle creators do not check whether their puzzle has one unique solution.

Puzzle

2	2
3	4

Solution 1

1	2	3	4
4	3	2	1
3	4	1	2
2	1	4	3

Solution 2

4	2	3	1
1	3	2	4
3	4	1	2
2	1	4	3

Some Interesting 9×9 Sudoku Facts

- How many 9×9 matrices deserve the title of Sudoku matrices?

$6,670,903,752,021,072,936,960 \approx 6.67 \times 10^{21}$

- What is the fewest number of givens that must be provided to create a 9×9 puzzle with a unique solution?

17; 35,396 distinct puzzles with 17 givens and a unique solution have been found. No unique solution puzzle with ≤ 16 givens has been found yet.

- Given one Sudoku matrix, could I make my own **Daily Sudoku Calendar**?

Puzzle		Unique Solution			Puzzle		Unique Solution	
2		1 2	3 4	$1 \Leftrightarrow 4$	2		4 2	3 1
	1	4 3	2 1			4	1 3	2 4
3		3 4	1 2		3		3 1	4 2
	4	2 1	4 3			1	2 4	1 3

By using mathematical operations 362,879 (≈ 991 years worth of) Sudoku matrices can be created from one 9×9 Sudoku matrix.

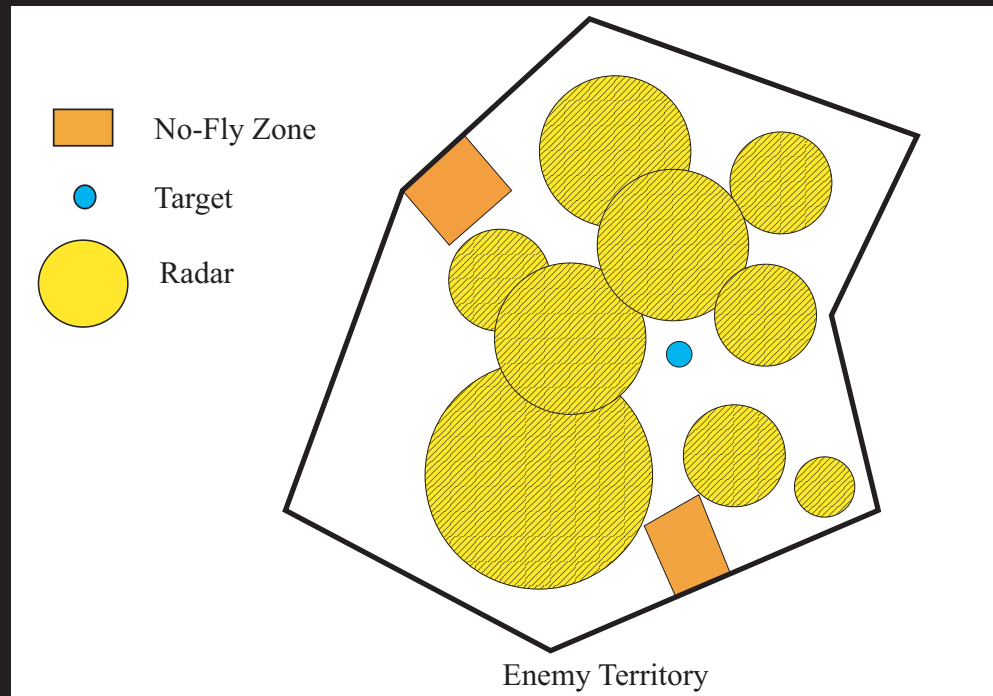
Military Applications

Outline

- Sudoku
- Military Applications
 - planning flight paths
 - disabling and herding communication in networks
- Ranking Applications
 - ranking on the World Wide Web
- Clustering and Data Mining Applications
 - clustering the Enron email dataset
 - clustering on terrorist networks

Flight Path Planning

(Lincoln Labs)



Objective:

create path that minimizes time over **radars**.

Constraints:

plane must fly over **target**

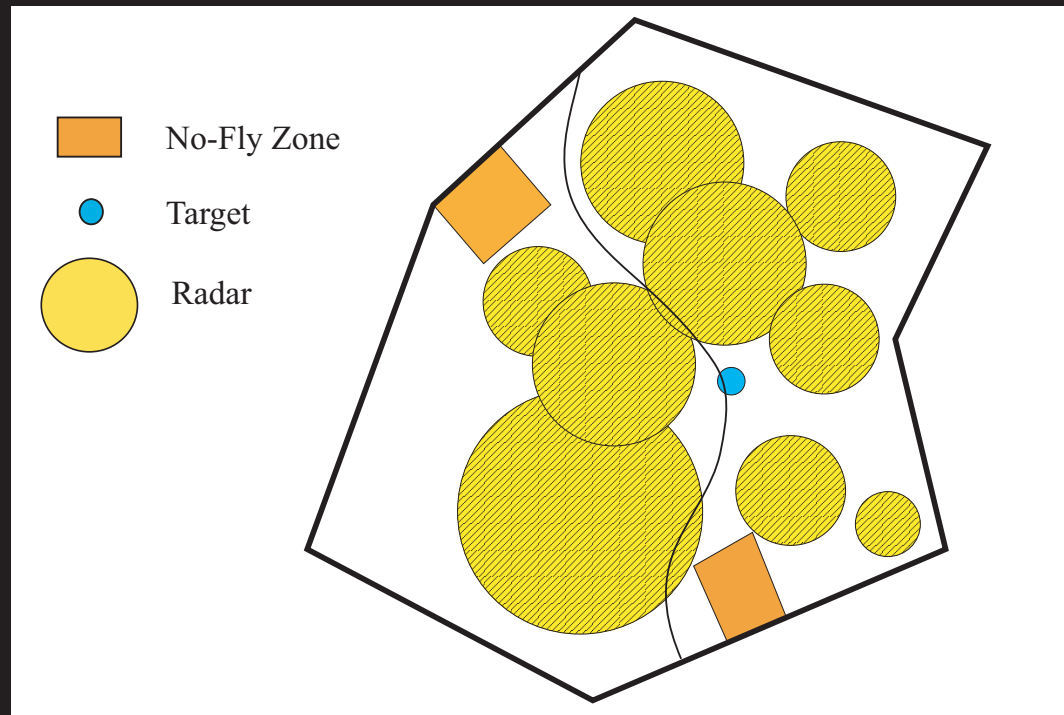
plane must avoid **no-fly zones**

plane cannot make unrealistic turns

plane has fixed amount of fuel

etc., etc., etc.

Flight Path Planning



Objective:

create path that minimizes time over **radars**.

Constraints:

plane must fly over **target**

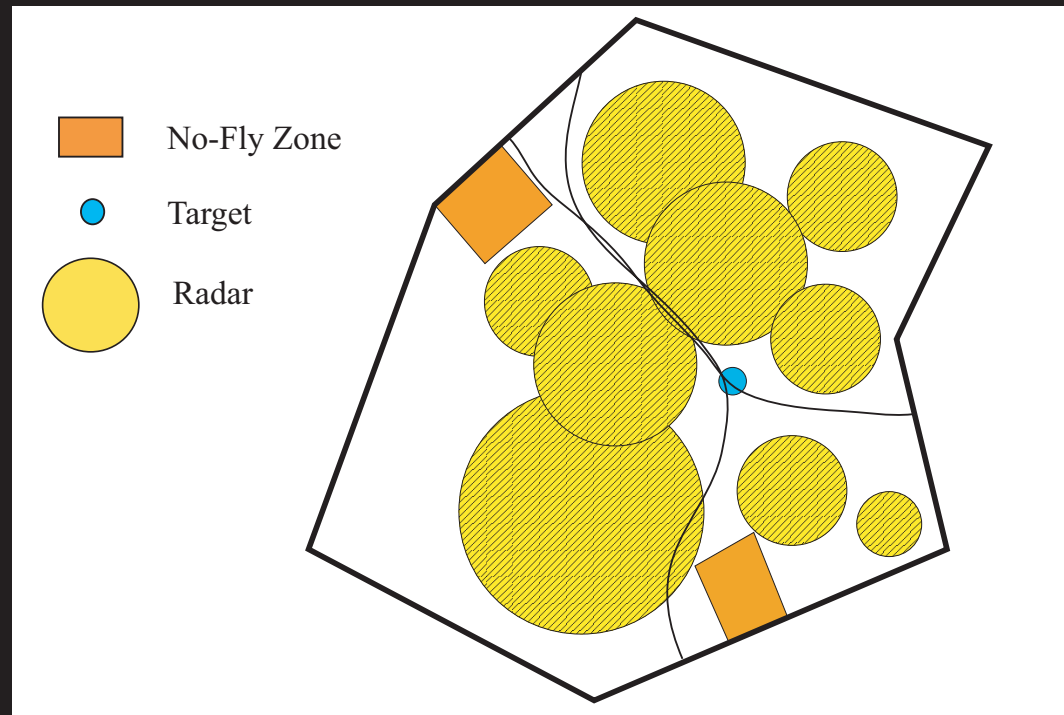
plane must avoid **no-fly zones**

plane cannot make unrealistic turns

plane has fixed amount of fuel

etc., etc., etc.

Flight Path Planning



Objective:

create path that minimizes time over **radars**.

Constraints:

plane must fly over **target**

plane must avoid **no-fly zones**

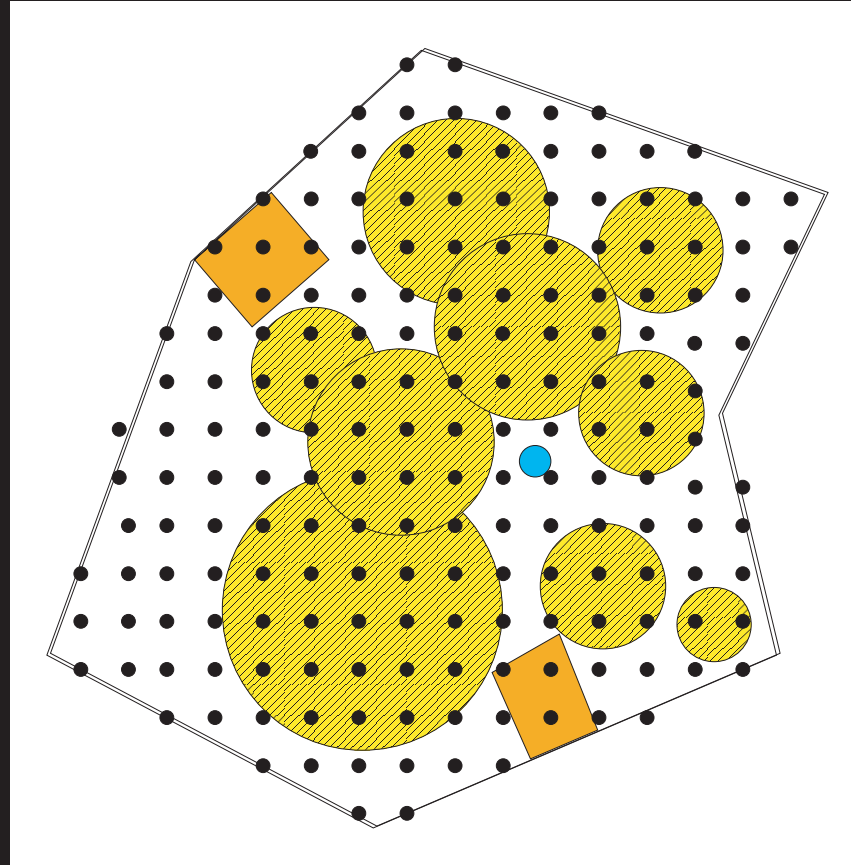
plane cannot make unrealistic turns

plane has fixed amount of fuel

etc., etc., etc.

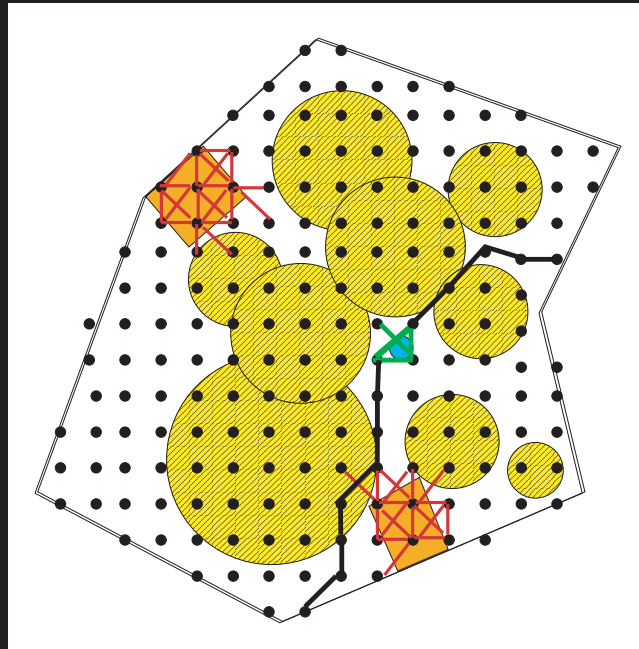
Flight Path Planning

Discretization

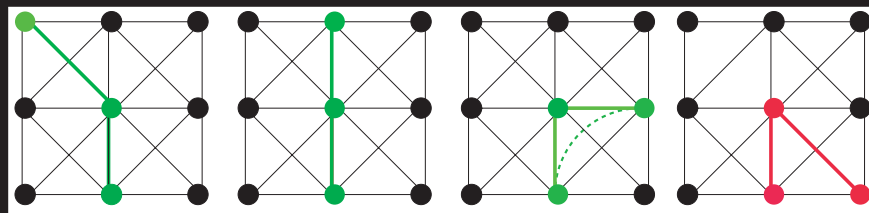


Flight Path Planning

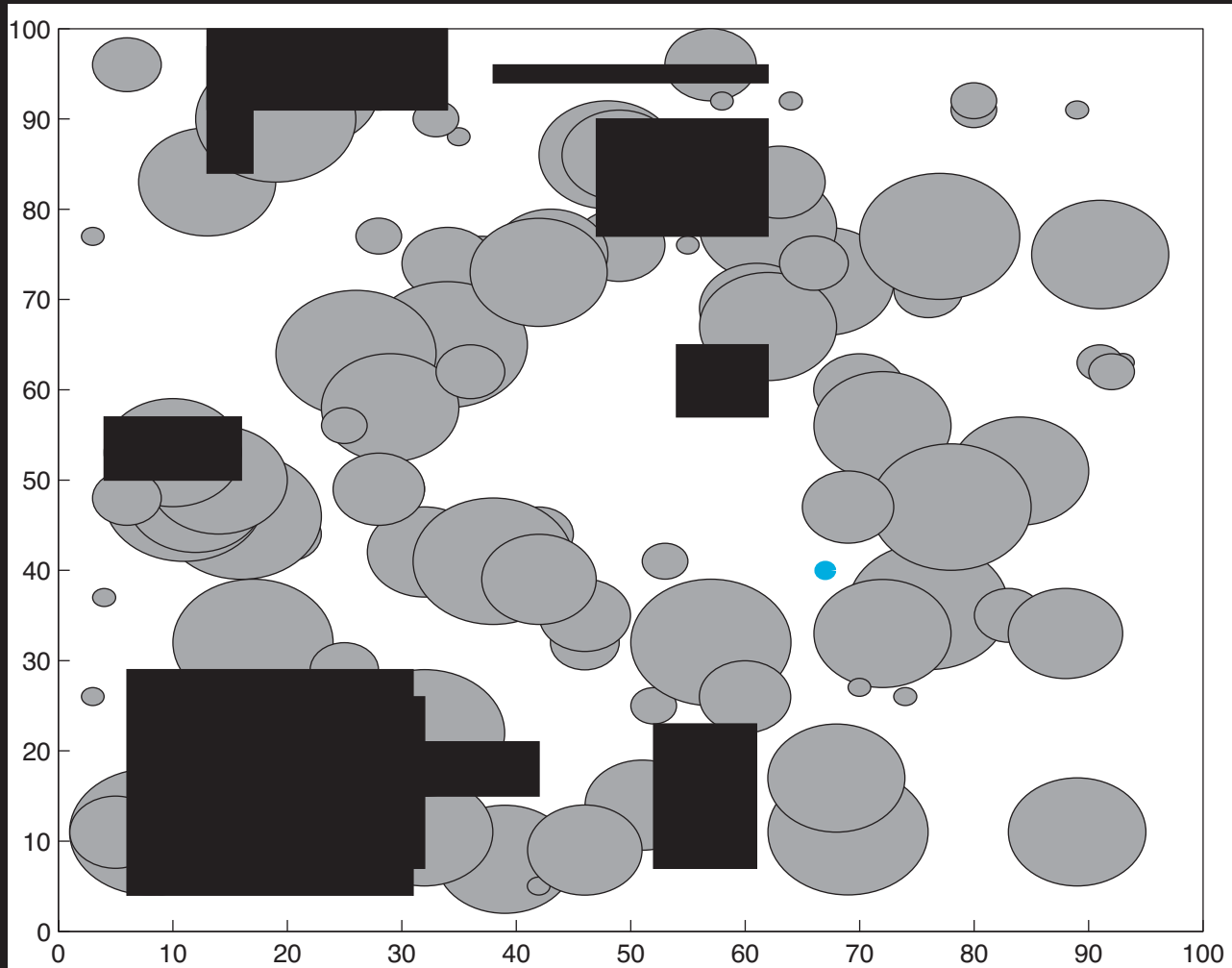
Connect the Dots



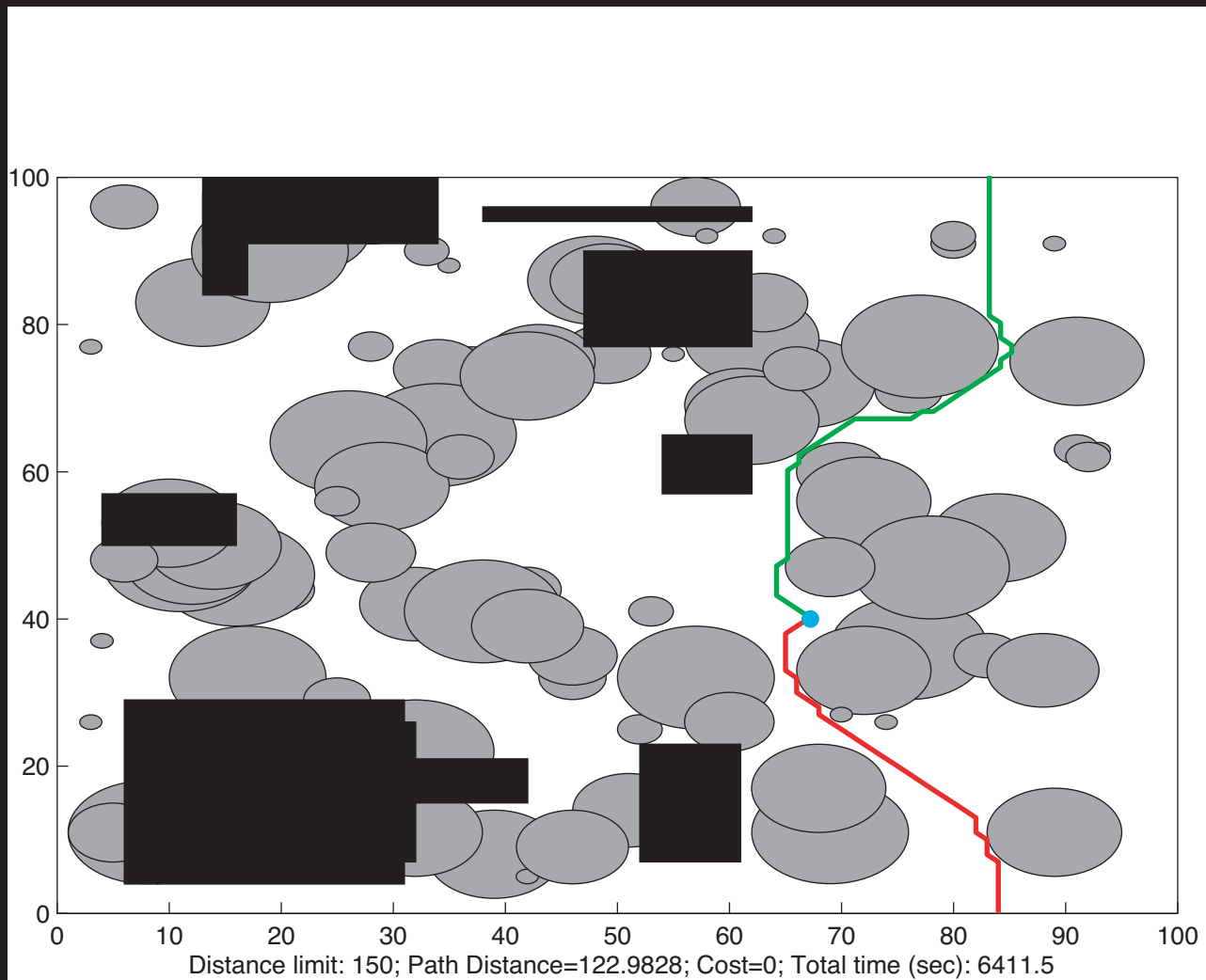
- plane must fly over target
- plane must avoid no-fly zones
- plane has fixed amount of fuel (total # path segments $\leq D$)
- plane cannot make unrealistic turns



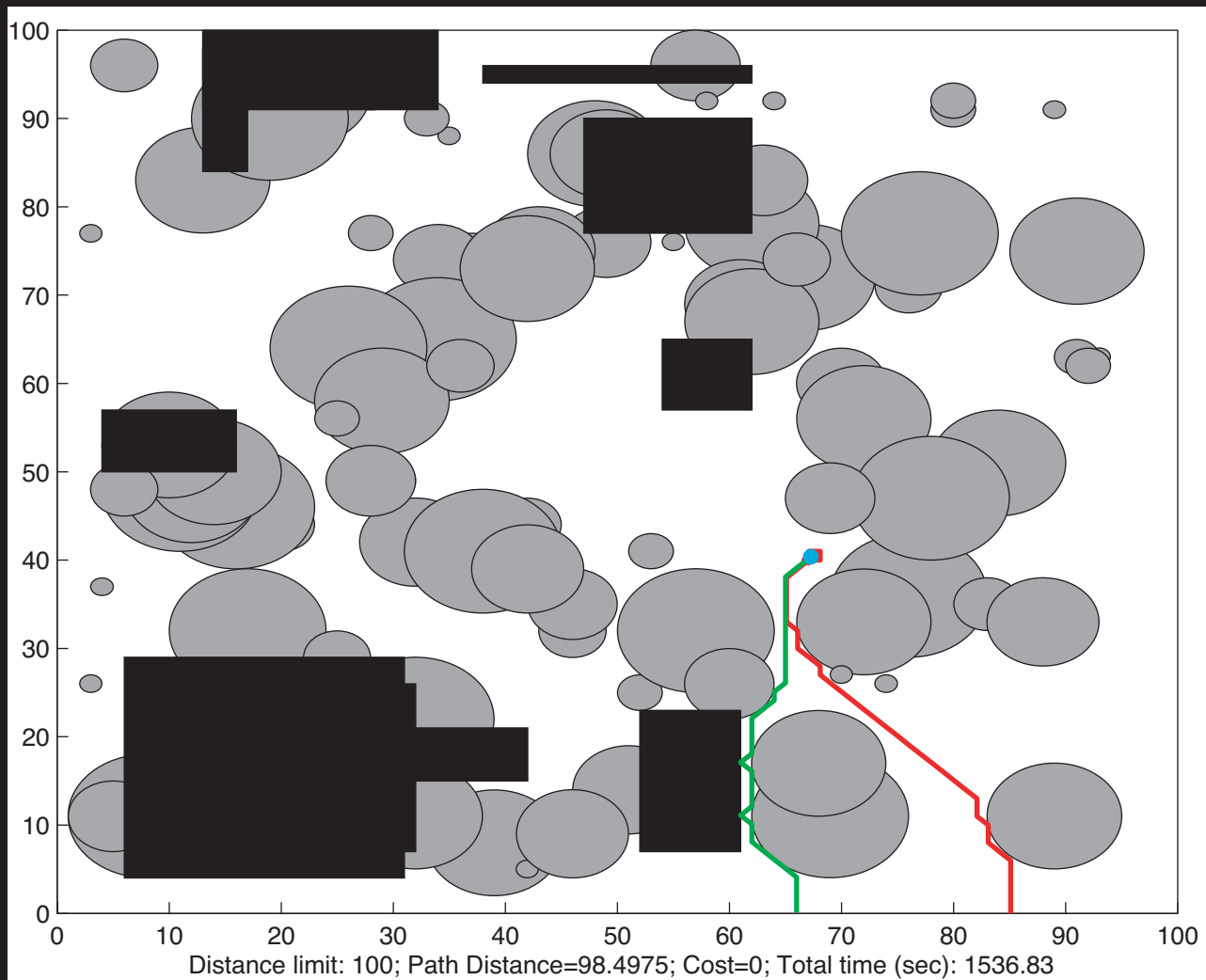
Flight Path Results



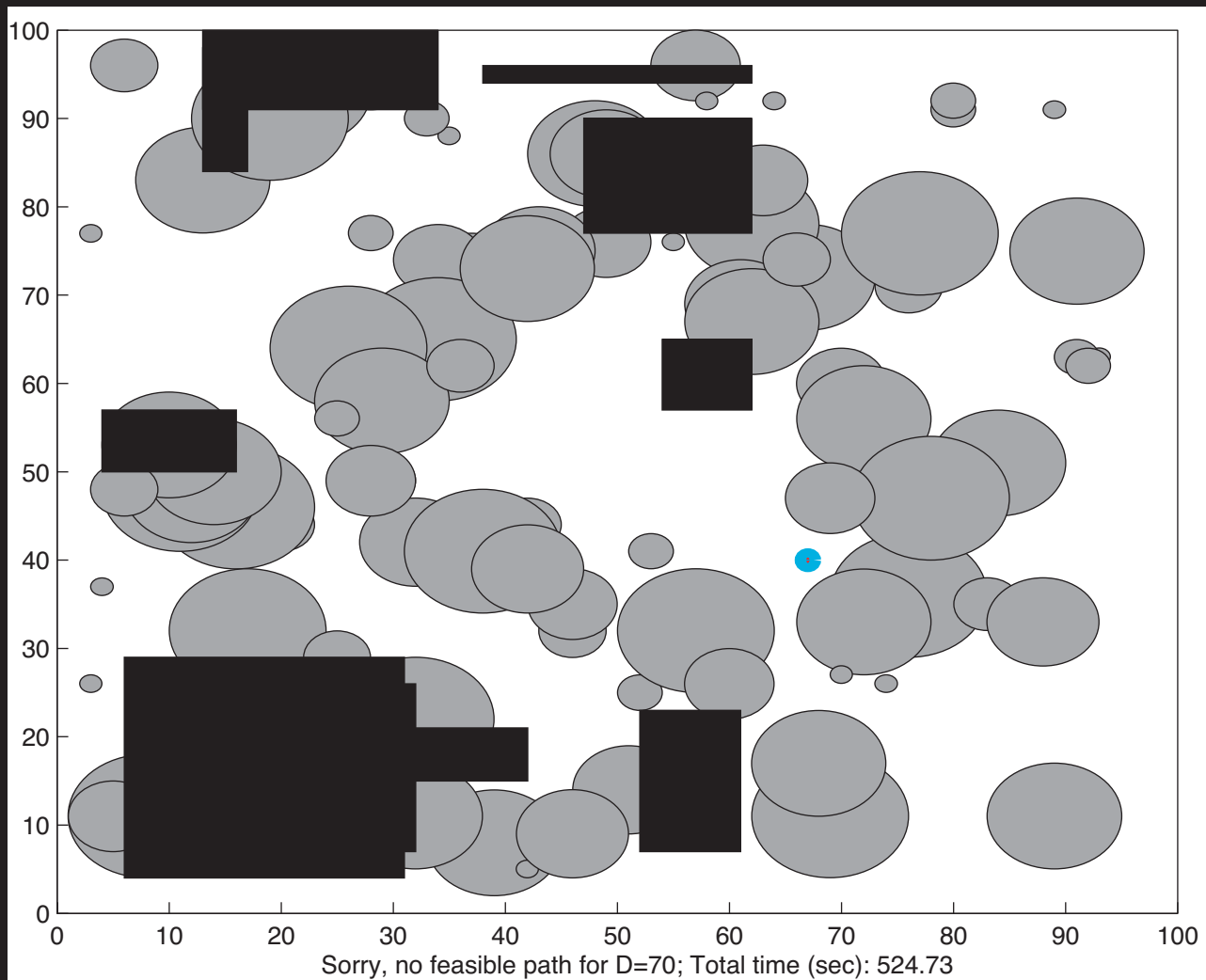
Flight Path Results



Flight Path Results



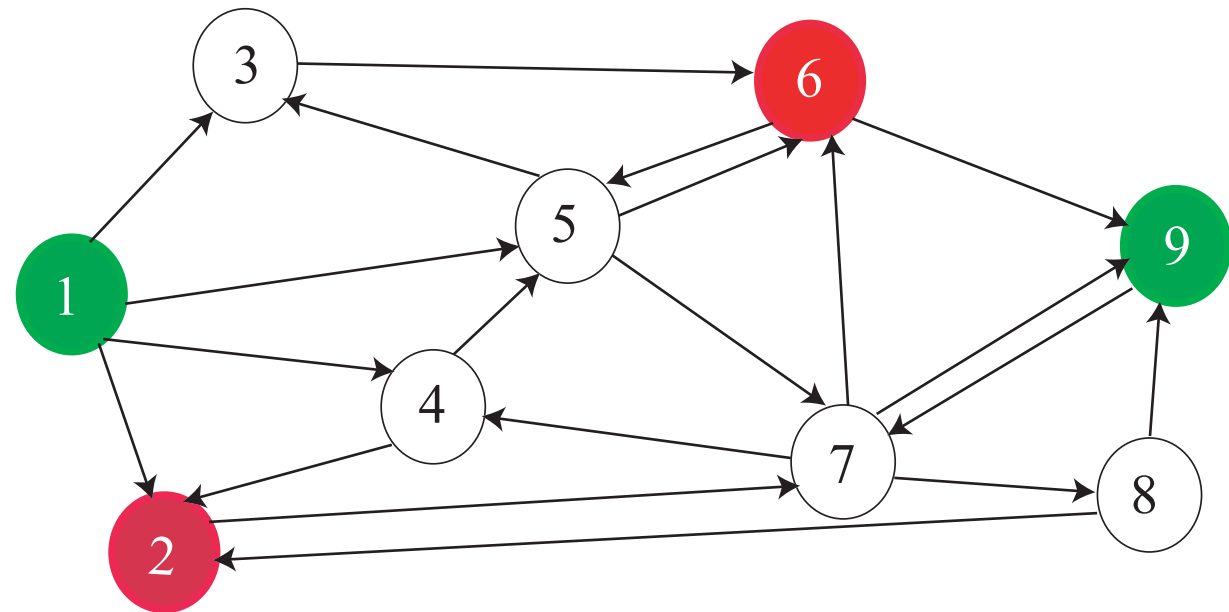
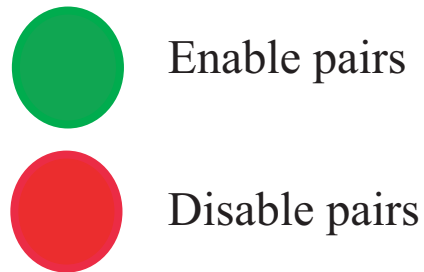
Flight Path Results



Outline

- Sudoku
- **Military Applications**
 - planning flight paths
 - **disabling and herding communication in networks**
- Ranking Applications
 - ranking on the World Wide Web
- Clustering and Data Mining Applications
 - clustering the Enron email dataset
 - clustering on terrorist networks

NSA Enemy Communication Networks



Objective:

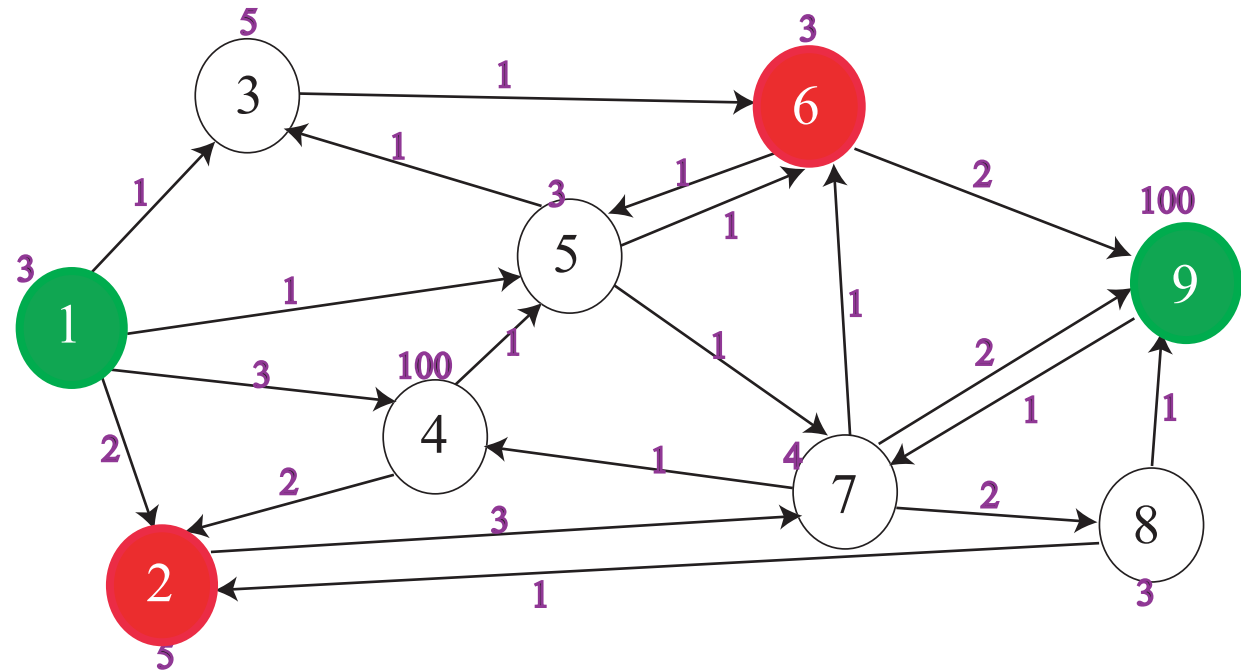
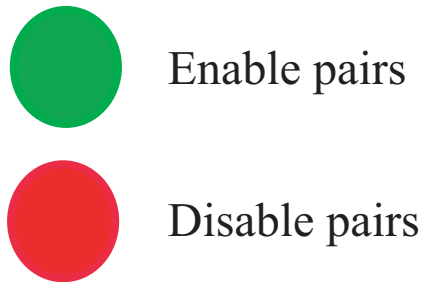
minimize cost associated with cutting links and nodes

Constraints:

enable communication between all **○—○** pairs

disable communication between all **○—○** pairs

NSA Communication Networks



Objective:

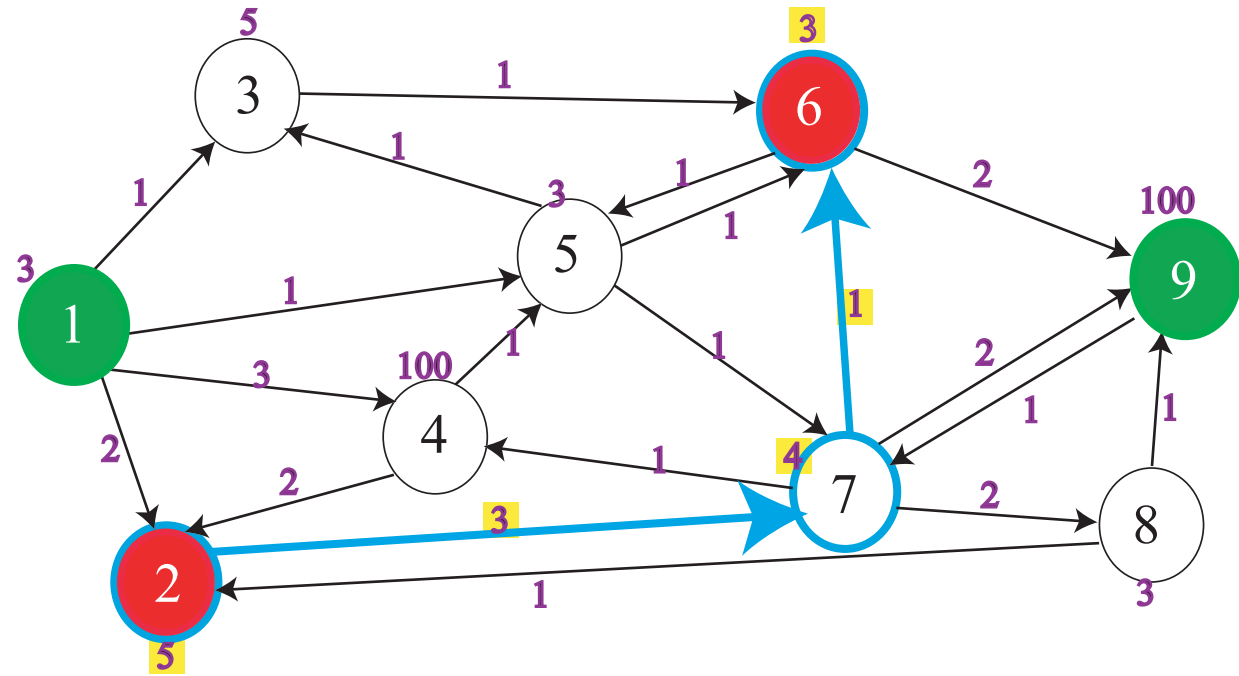
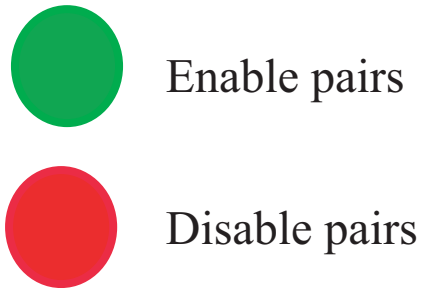
minimize **cost** associated with cutting links and nodes

Constraints:

enable communication between all **○—○** pairs

disable communication between all **○—○** pairs

NSA Communication Networks



Objective:

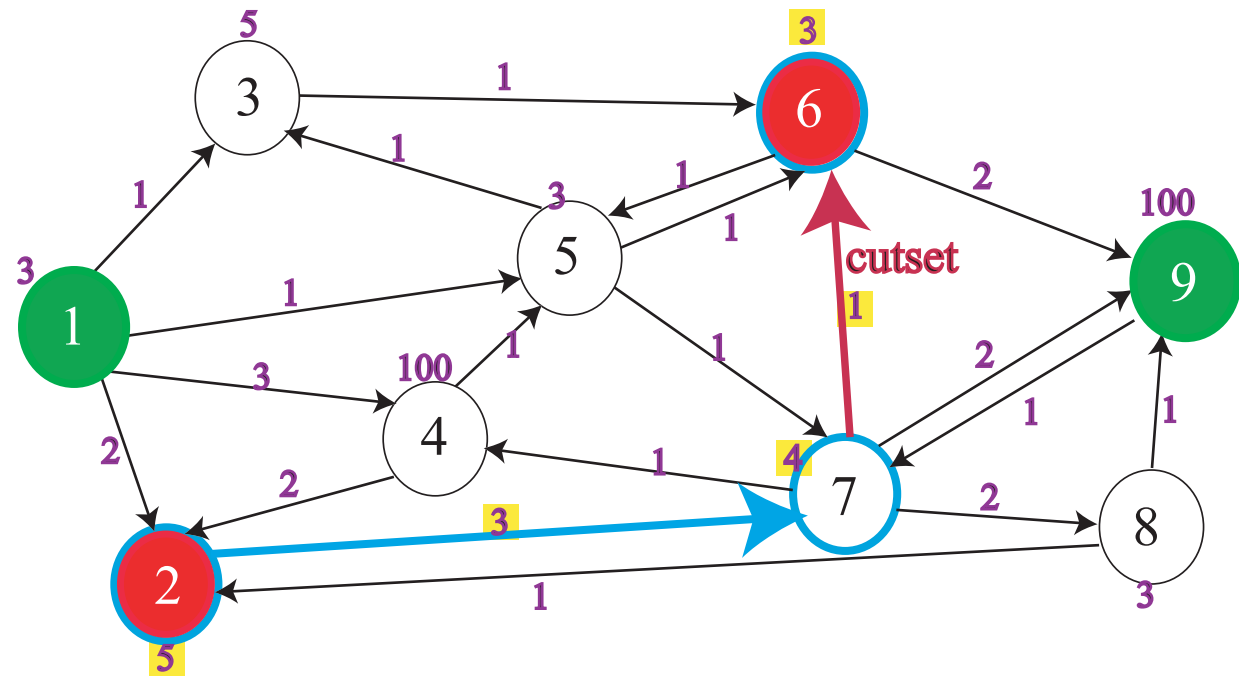
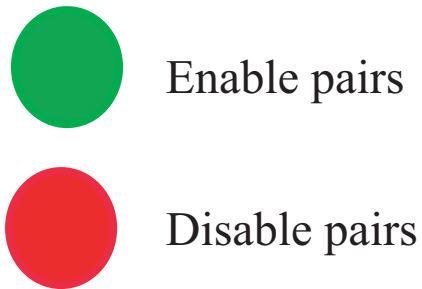
minimize **cost** associated with cutting links and nodes

Constraints:

enable communication between all **○—○** pairs

disable communication between all **○—○** pairs

NSA Communication Networks



Objective:

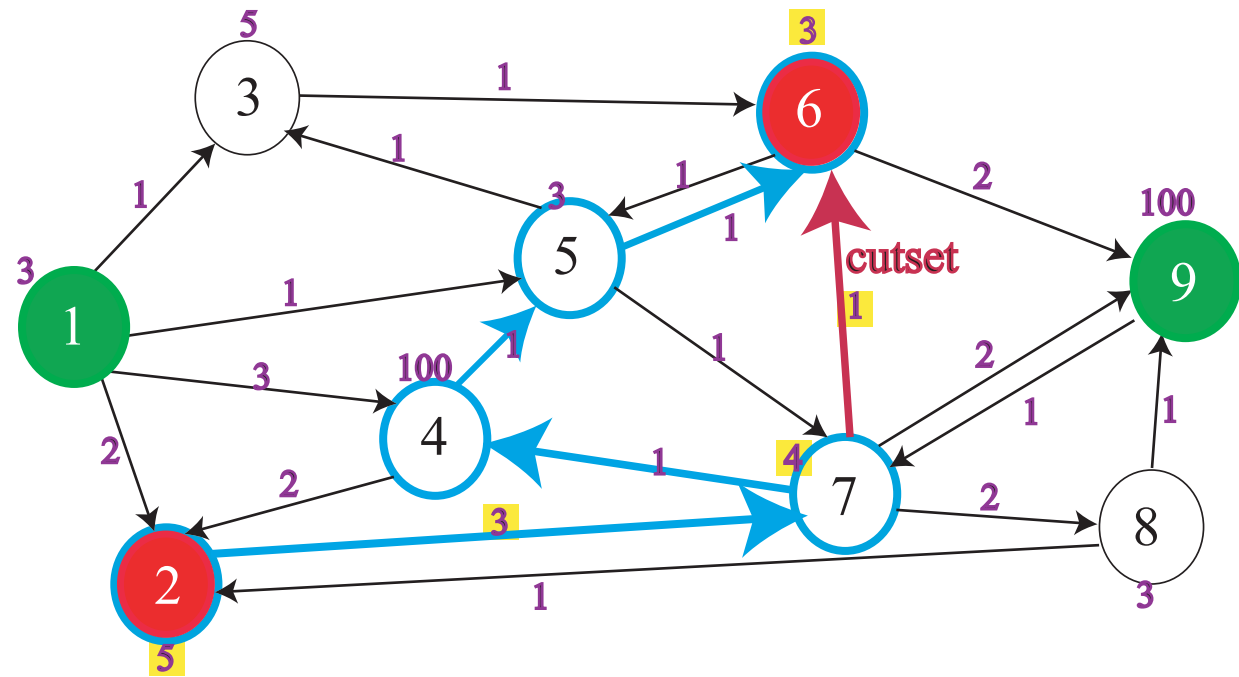
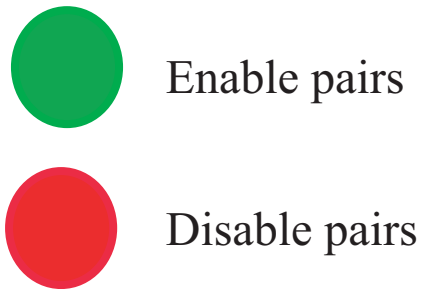
minimize cost associated with cutting links and nodes

Constraints:

enable communication between all O-O pairs

disable communication between all O-O pairs

NSA Communication Networks



Objective:

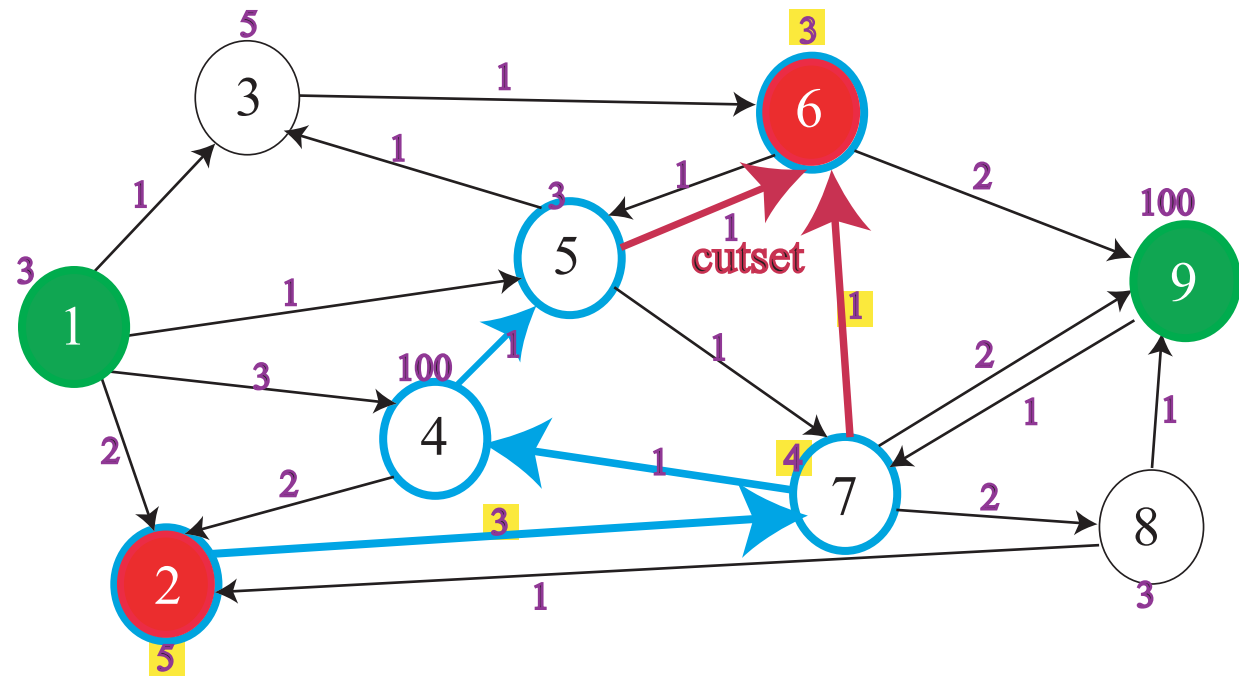
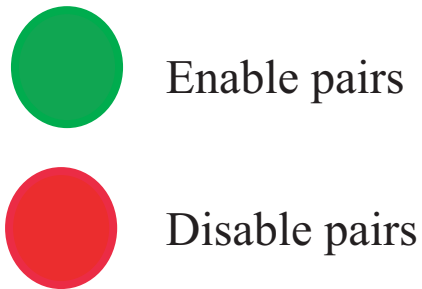
minimize cost associated with cutting links and nodes

Constraints:

enable communication between all O-O pairs

disable communication between all O-O pairs

NSA Communication Networks



Objective:

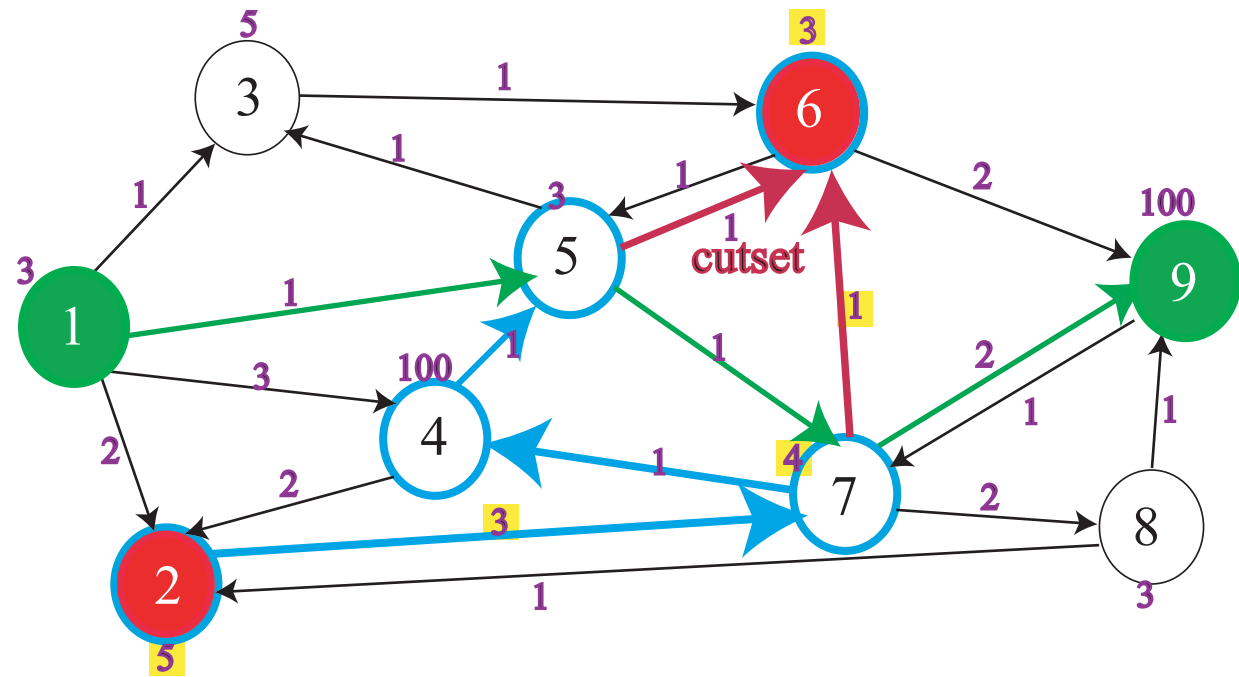
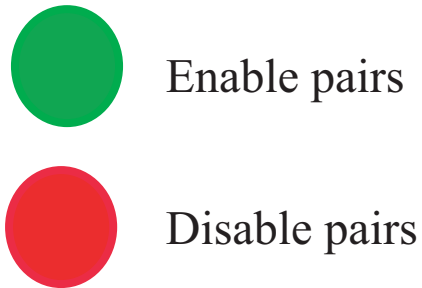
minimize cost associated with cutting links and nodes

Constraints:

enable communication between all \bigcirc – \bigcirc pairs

disable communication between all \bigcirc – \bigcirc pairs

NSA Communication Networks



Objective:

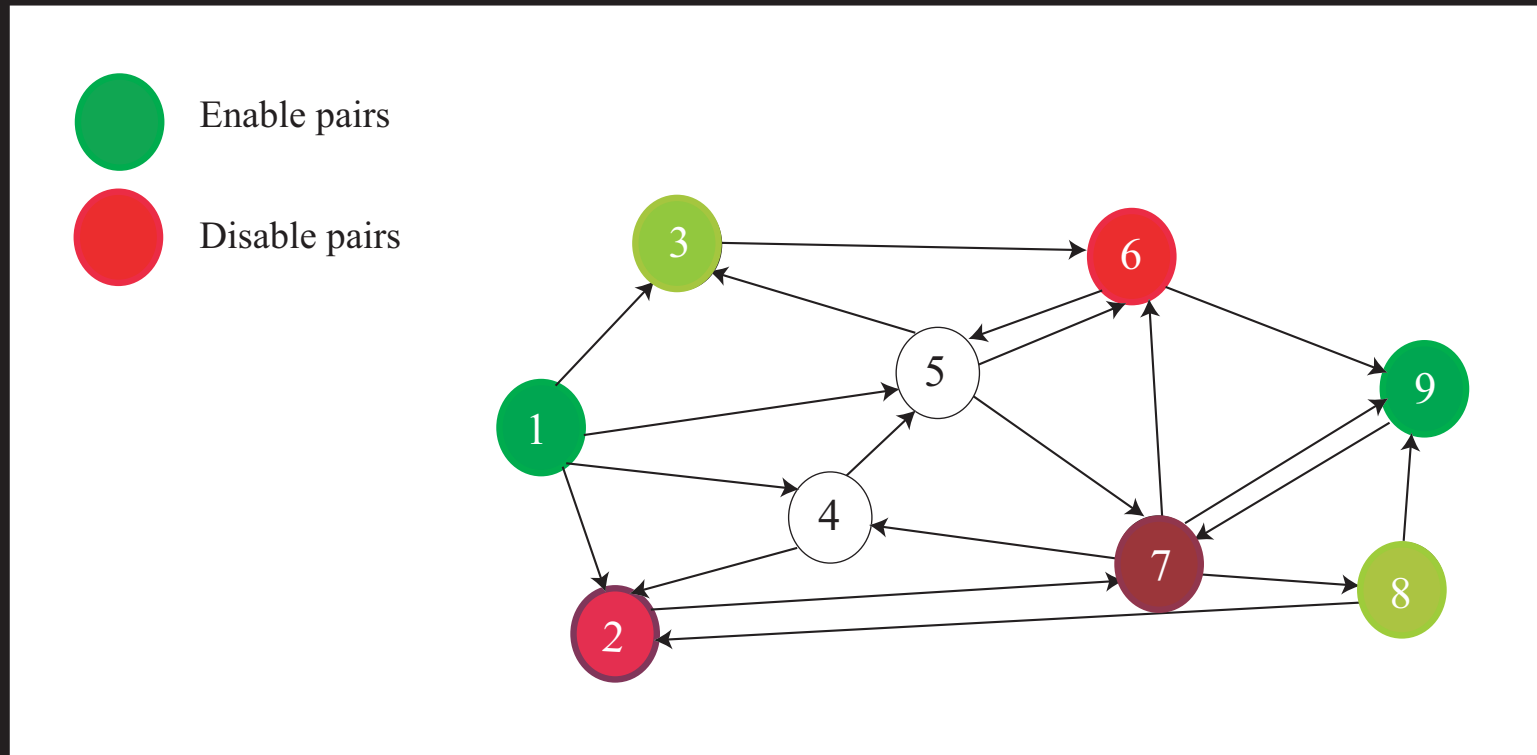
minimize cost associated with cutting links and nodes

Constraints:

enable communication between all \bigcirc – \bigcirc pairs

disable communication between all \bigcirc – \bigcirc pairs

Multiple enable-disable pairs



Objective:

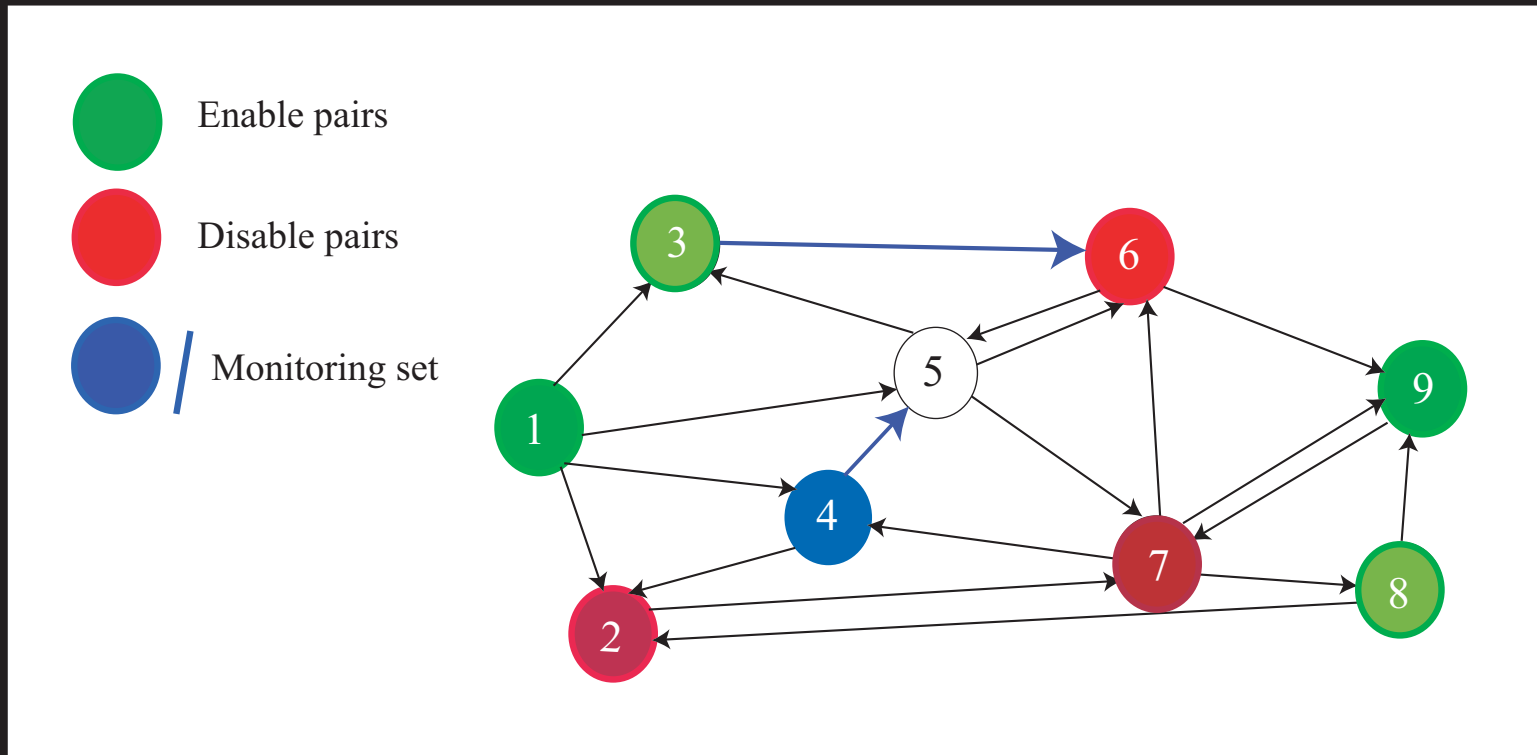
minimize cost associated with cutting links and nodes

Constraints:

enable communication between all **○—○** pairs

disable communication between all **○—○** pairs

Herding Problem



Objective:

minimize cost associated with cutting links and nodes

Constraints:

enable communication between all **○—○** pairs

disable communication between all **○—○** pairs

herd all communication over monitored set

Ranking Applications

Outline

- Sudoku
- Military Applications
 - planning flight paths
 - disabling and herding communication in networks
- Ranking Applications
 - ranking on the World Wide Web
- Clustering and Data Mining Applications
 - clustering the Enron email dataset
 - clustering on terrorist networks

the pre-1998 Web

Yahoo

- hierarchies of sites
- organized by humans

Best Search Techniques

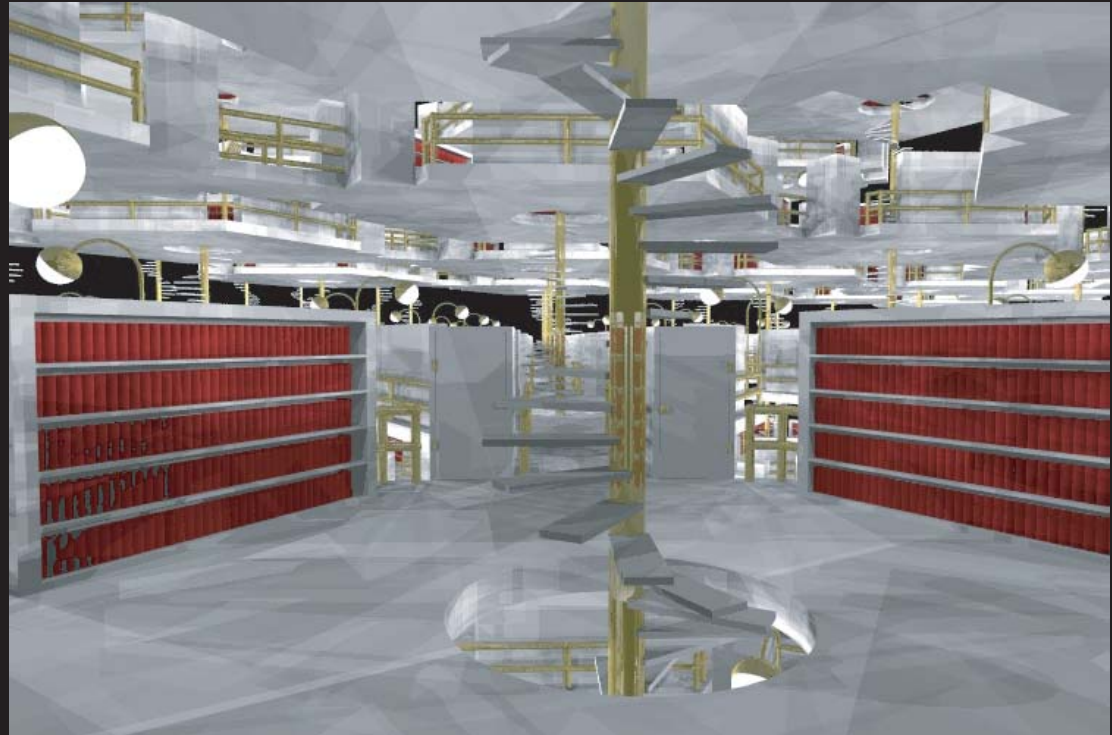
- word of mouth
- expert advice

Overall Feeling of Users

- Jorge Luis Borges' 1941 short story, *The Library of Babel*

When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. All men felt themselves to be the masters of an intact and secret treasure. There was no personal or world problem whose eloquent solution did not exist in some hexagon.

... As was natural, this inordinate hope was followed by an excessive depression. The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible, seemed almost intolerable.



1998 ... enter Link Analysis

Change in User Attitudes about Web Search

Today

- “It’s not my homepage, but it might as well be. I use it to ego-surf. I use it to read the news. Anytime I want to find out anything, I use it.” - **Matt Groening, creator and executive producer, The Simpsons**
- “I can’t imagine life without Google News. Thousands of sources from around the world ensure anyone with an Internet connection can stay informed. The diversity of viewpoints available is staggering.” - **Michael Powell, chair, Federal Communications Commission**
- “Google is my rapid-response research assistant. On the run-up to a deadline, I may use it to check the spelling of a foreign name, to acquire an image of a particular piece of military hardware, to find the exact quote of a public figure, check a stat, translate a phrase, or research the background of a particular corporation. It’s the Swiss Army knife of information retrieval.” - **Garry Trudeau, cartoonist and creator, Doonesbury**

Ranking on the Web

the pre-1998 Web

⋮

- border patrol: 4; 567; 809; 1103;

⋮

- hezbollah: 9; 12; 339; 942; 15158;

⋮

- global warming: 178; 12980; 445532;

Index

- k -step transition matrix, 179
- a** vector, 37, 38, 75, 80
- A9, 142
- absolute error, 104
- absorbing Markov chains, 185
- absorbing states, 185
- accuracy, 79–80
- adaptive PageRank method, 89–90
- Adar, Eytan, 146
- adjacency list, 77
- adjacency matrix, 33, 76, 116, 132, 169
- advertising, 45
- aggregated chain, 197
- aggregated chains, 195
- aggregated transition matrix, 105
- aggregated transition probability, 197
- aggregation, 94–97
 - approximate, 102–104
 - exact, 104–105
 - exact vs. approximate, 105–107
 - iterative, 107–109
 - partition, 109–112
- aggregation in Markov chains, 197
- aggregation theorem, 105
- Aitken extrapolation, 91
- Alexa traffic ranking, 138
- algebraic multiplicity, 157
- algorithm
 - PageRank, 40
 - Aitken extrapolation, 92
 - dangling node PageRank, 82, 83
 - HITS, 116
 - iterative aggregation updating, 108
 - personalized PageRank power method, 49
 - quadratic extrapolation, 93
 - query-independent HITS, 124
- α parameter, 37, 38, 41, 47–48
- Amazon’s traffic rank, 142
- anchor text, 48, 54, 201
- Ando, Albert, 110
- aperiodic, 36, 133
- aperiodic Markov chain, 176
- Application Programming Interface (API), 65, 73, 97
- approximate aggregation, 102–104
- arc, 201
- Arrow, Kenneth, 136
- asymptotic convergence rate, 165
- asymptotic rate of convergence, 41, 47, 101, 119, 125
- Atlas of Cyberspace*, 27
- authority, 29, 201
- authority Markov chain, 132
- authority matrix, 117, 201
- authority score, 115, 201
- authority vector, 201
- Babbage, Charles, 75
- back button, 84–86
- BadRank, 141
- Barabasi, Albert-Laszlo, 30
- Berry, Michael, 7
- bibliometrics, 32, 123
- bipartite undirected graph, 131
- BlockRank, 94–97, 102
- blog, 55, 144–146, 201
- Boldi, Paolo, 79
- Boolean model, 5–6, 201
- bounce back, 84–86
- bowtie structure, 134
- Brezinski, Claude, 92
- Brin, Sergey, 25, 205
- Browne, Murray, 7
- Bush, Vannevar, 3, 10
- Campbell, Lord John, 23
- canonical form, reducible matrix, 182
- censored chain, 104
- censored chains, 194
- censored distribution, 104, 195
- censored Markov chain, 194
- ensorship, 146–147
- Cesàro sequence, 162
- Cesàro summability, stochastic matrix, 182
- characteristic polynomial, 120, 156
- Chebyshev extrapolation, 92
- Chien, Steve, 102
- cloaking, 44
- clustering search results, 142–143
- co-citation, 123, 201
- co-reference, 123, 201
- Collatz–Wielandt formula, 168, 172
- complex networks, 30
- compressed matrix storage, 76
- condition number, 59, 71, 155
- Condorcet, 136
- connected components, 127, 133

Ranking on the Web

the pre-1998 Web

⋮

- border patrol: 4; 567; 809; 1103; . . . (8,700,000 in total)

⋮

- hezbollah: 9; 12; 339; 942; 15158; . . . (15,100,000 in total)

⋮

- global warming: 178; 12980; 445532; . . . (33,200,000 in total)

Ranking on the Web

the pre-1998 Web

⋮

- border patrol: 4; 567; 809; 1103; . . . (8,700,000 in total)

⋮

- hezbollah: 9; 12; 339; 942; 15158; . . . (15,100,000 in total)

⋮

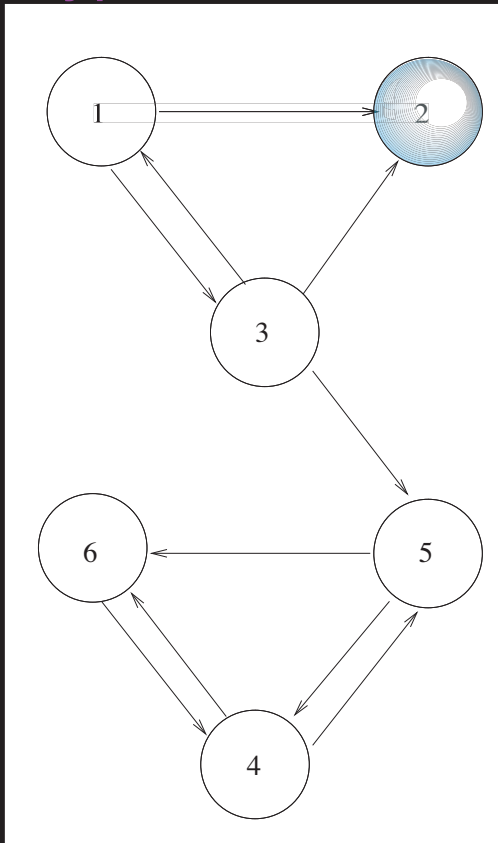
- global warming: 178; 12980; 445532; . . . (33,200,000 in total)

too many results per search term

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote

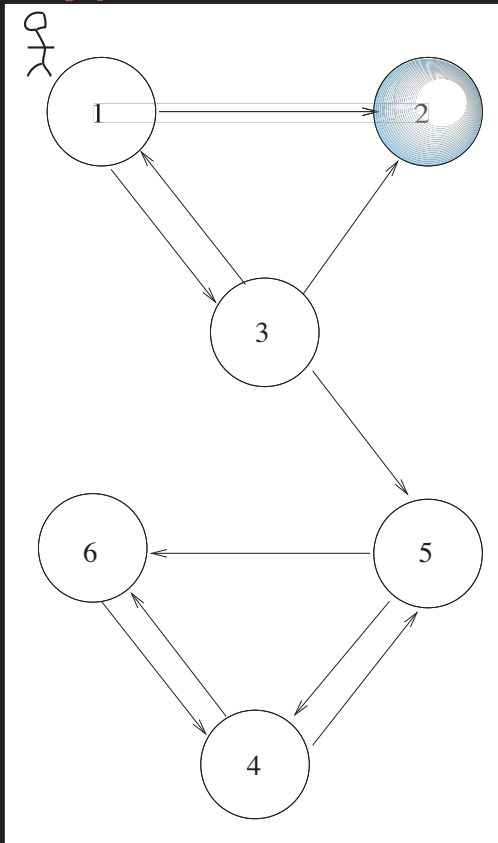


Markov chain

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

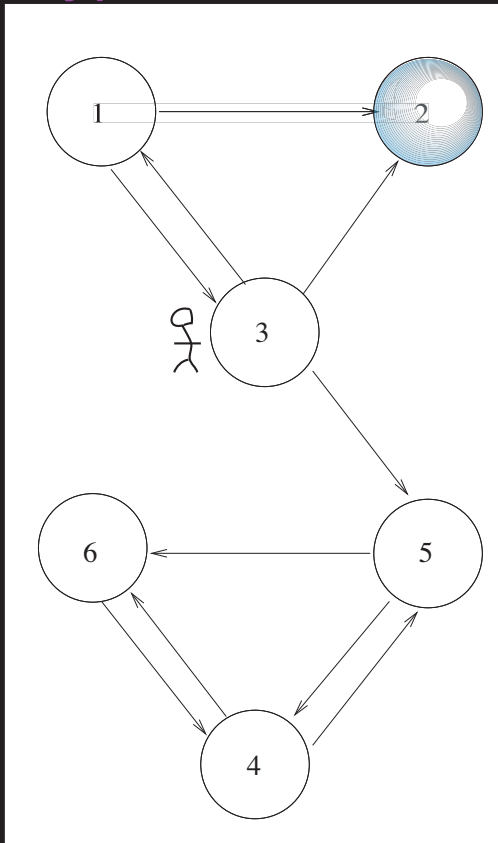
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

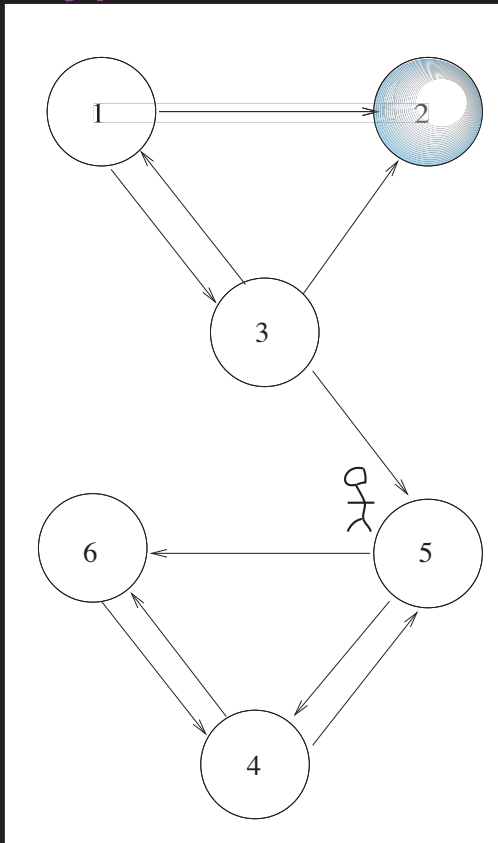
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

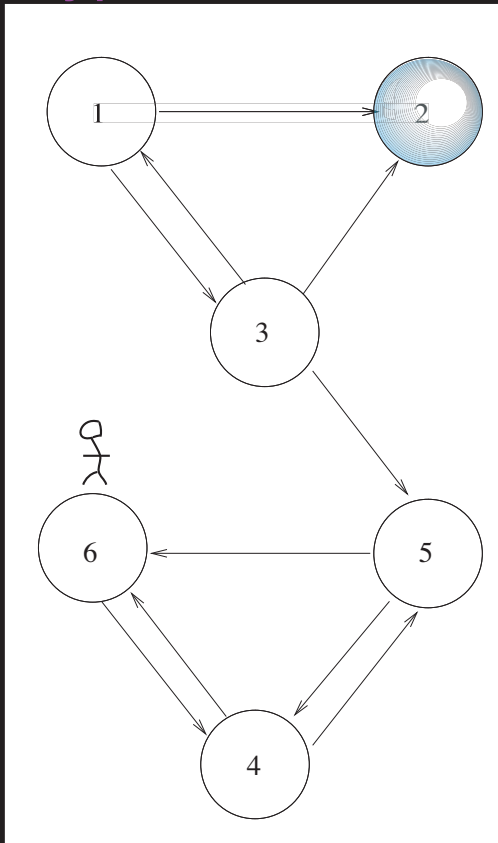
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

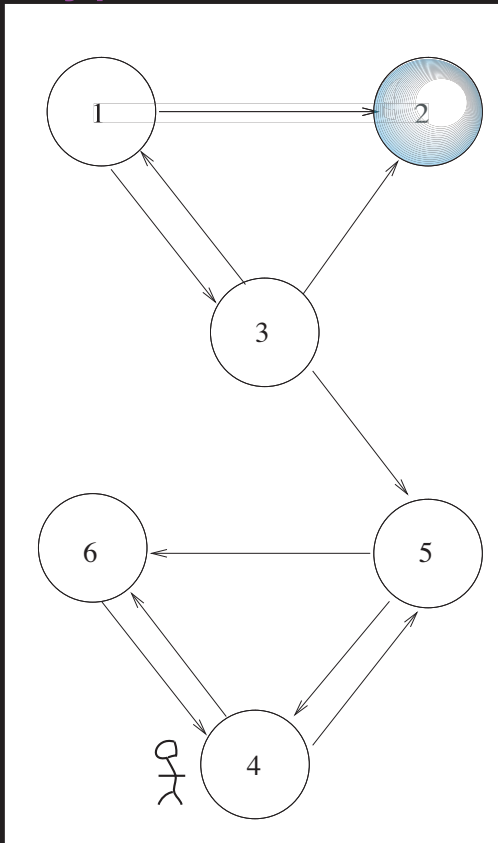
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

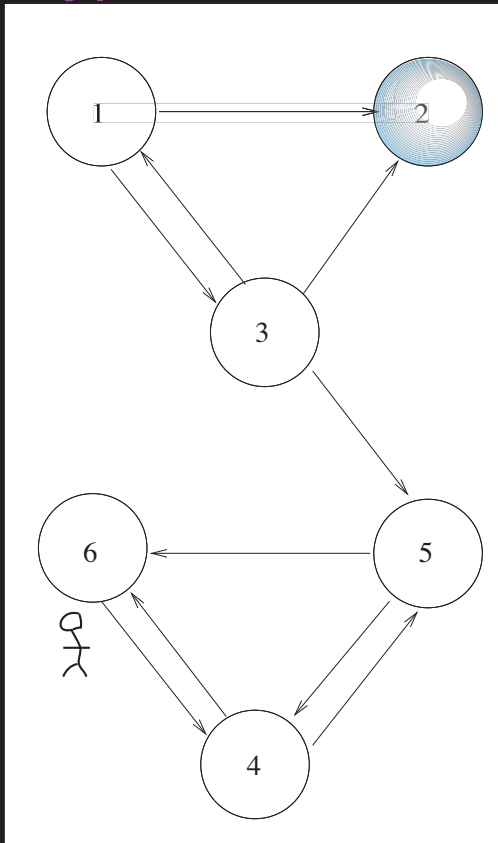
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

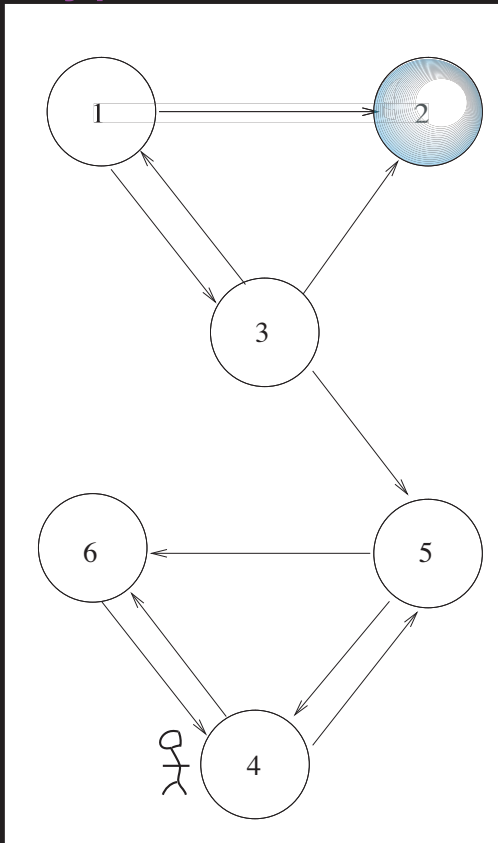
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

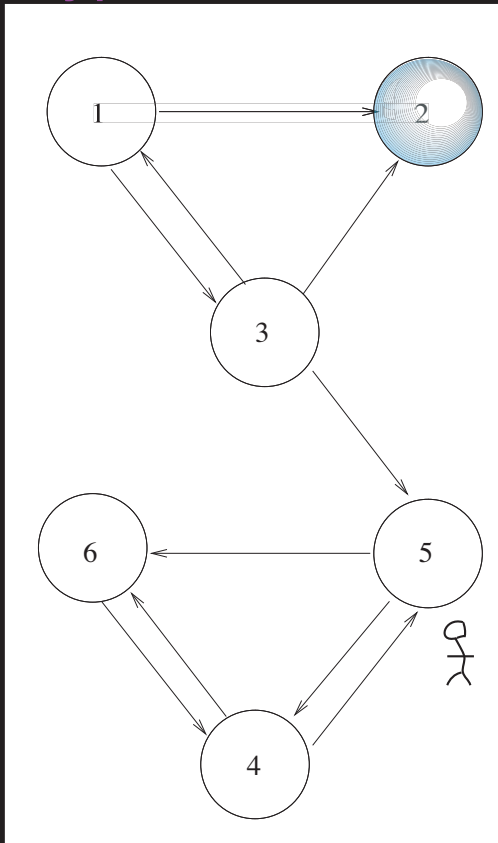
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

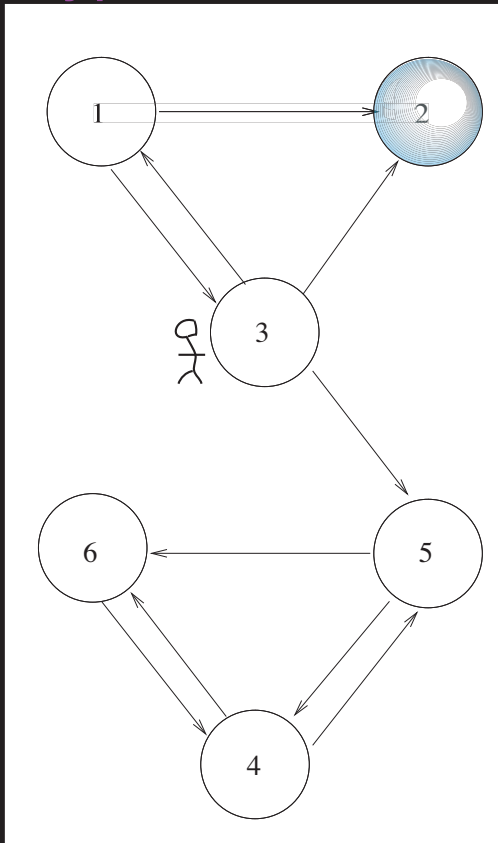
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

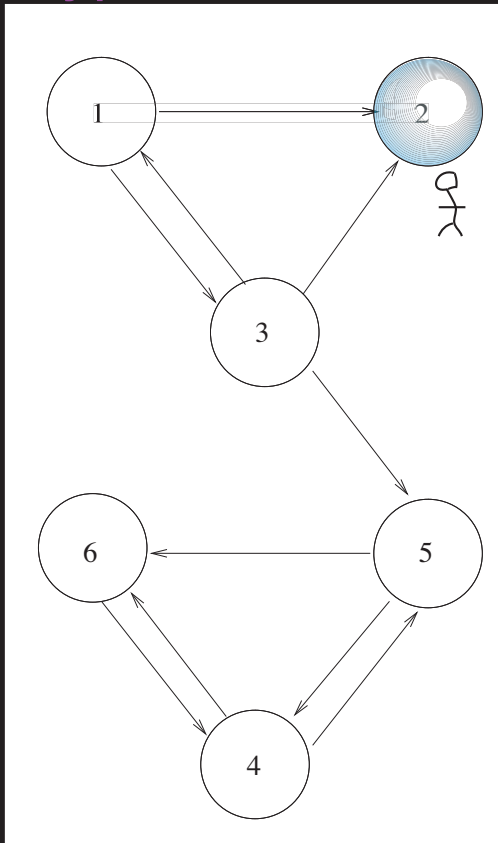
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote

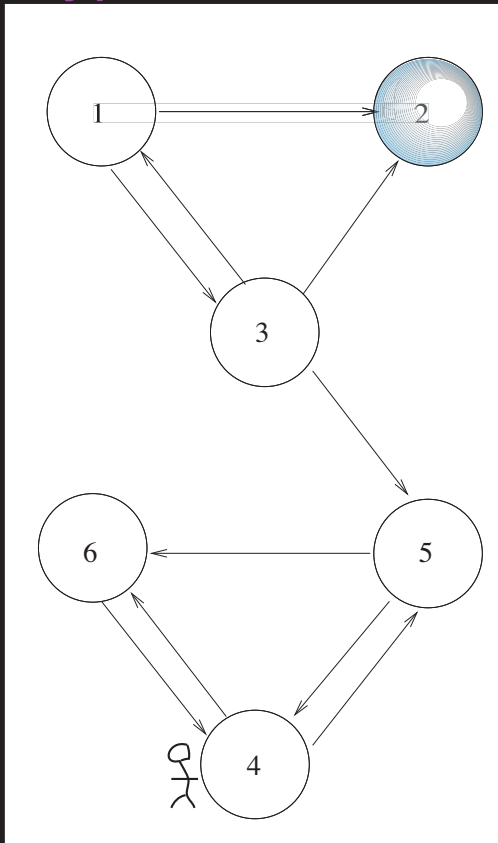


page 2 is a dangling node

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote



surfer “teleports”

Ranking with a Random Surfer

- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or *proportion of time*, the surfer spends on each page to create ranking of webpages.

Ranking with a Random Surfer

- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or *proportion of time*, the surfer spends on each page to create ranking of webpages.

Proportion of Time

Page 1 = .04

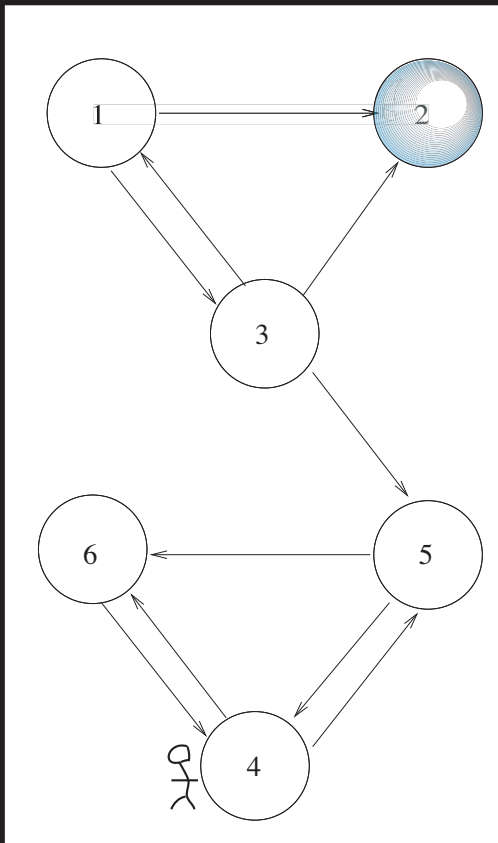
Page 2 = .05

Page 3 = .04

Page 4 = .38

Page 5 = .20

Page 6 = .29



Ranked List of Pages

Page 4

Page 6

Page 5

Page 2

Page 1

Page 3

Clustering and Data Mining Applications

Outline

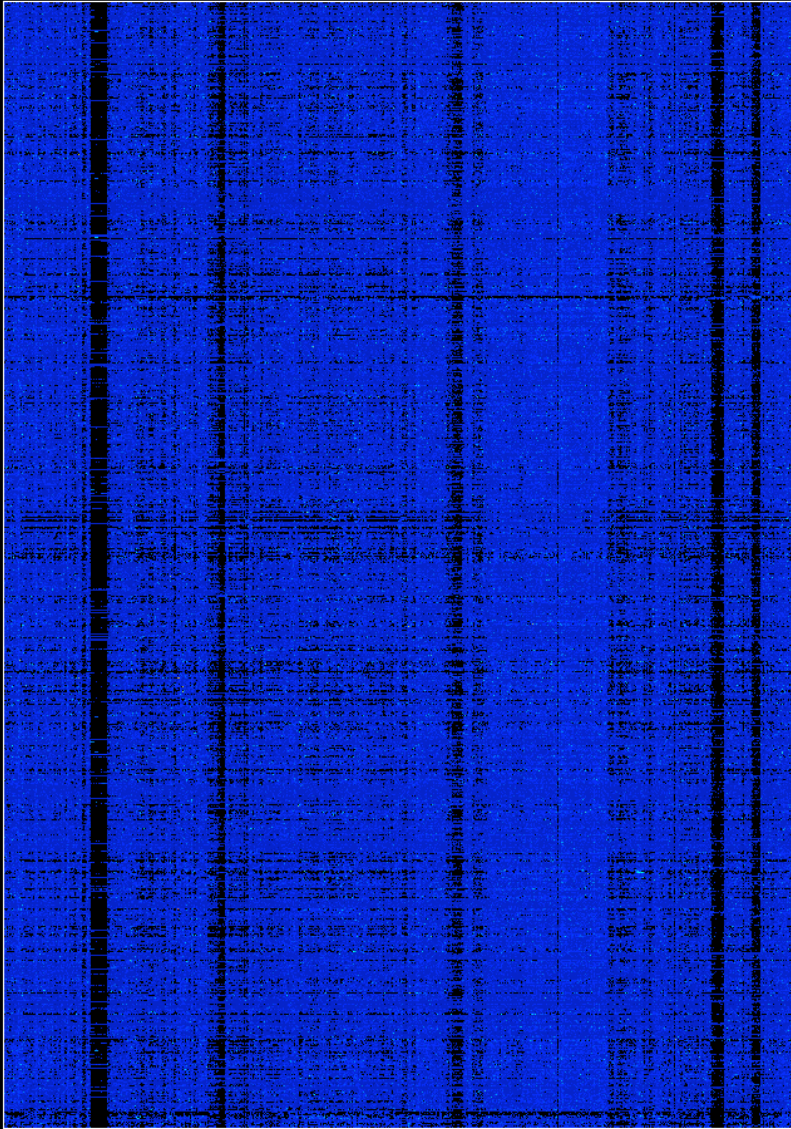
- Sudoku
- Military Applications
 - planning flight paths
 - disabling and herding communication in networks
- Ranking Applications
 - ranking on the World Wide Web
- Clustering and Data Mining Applications
 - clustering the Enron email dataset
 - clustering on terrorist networks

The Enron Email Dataset

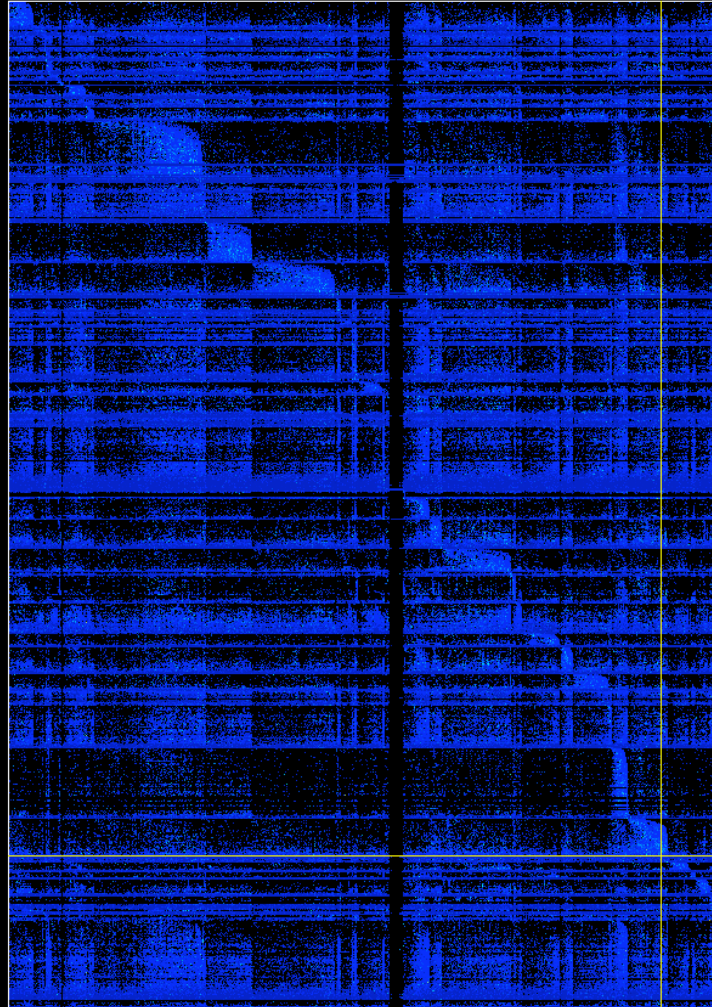
(SAS)

- PRIVATE email collection of 150 Enron employees during 2001
- 92,000 terms and 65,000 messages
- Term-by-Message Matrix

$$\begin{array}{c} \vdots \\ subpoena \\ dynegy \\ \vdots \end{array} \begin{pmatrix} fastow1 & fastow2 & skilling1 & \dots \\ \vdots & \vdots & \vdots & \dots \\ 2 & 0 & 1 & \dots \\ 0 & 3 & 0 & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix}$$



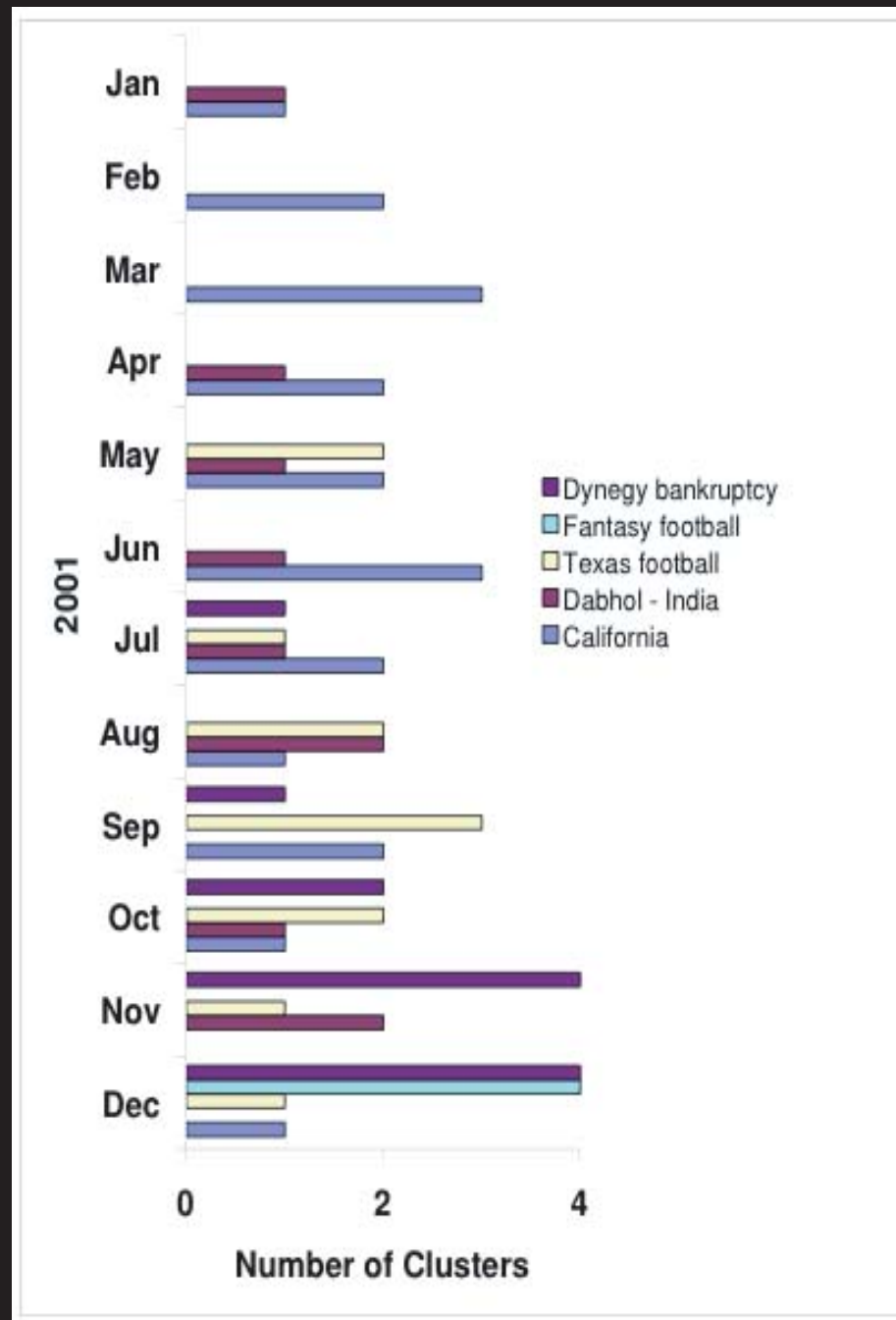
```
row [78198]: destroy
col [59746]: sanders-rsempra8
val [78198][59746]: 0.459000
```



Clustering the Enron Email Dataset

Feature Index (k)	Cluster Size	Topic Description	Dominant Terms
10	497	California	ca, cpuc , gov, socalgas , sempra, org, sce, gmssr, aelaw, ci
23	43	Louise Kitchen named top woman by Fortune	evp, fortune , britain, woman, ceo , avon, fiorinai, cfo, hewlett, packard
26	231	Fantasy football	game, wr, qb, play, rb, season, injury, updated, fantasy, image
33	233	Texas longhorn football newsletter	UT, orange, longhorn[s], texas, true, truorange, recruiting, oklahoma defensive
34	65	Enron collapse	partnership[s] , fastow , shares, sec , stock, shareholder, investors, equity, lay
39	235	Emails about India	dahhol , dpc , india , mseb , maharashtra , indian, lenders, delhi, foreign, minister
46	127	Enron collapse	dow, debt, reserved, wall, copyright jones, cents, analysts, reuters, spokesman

Tracking Enron clusters over time



Visualizing Clusters in the **Enron Dataset**



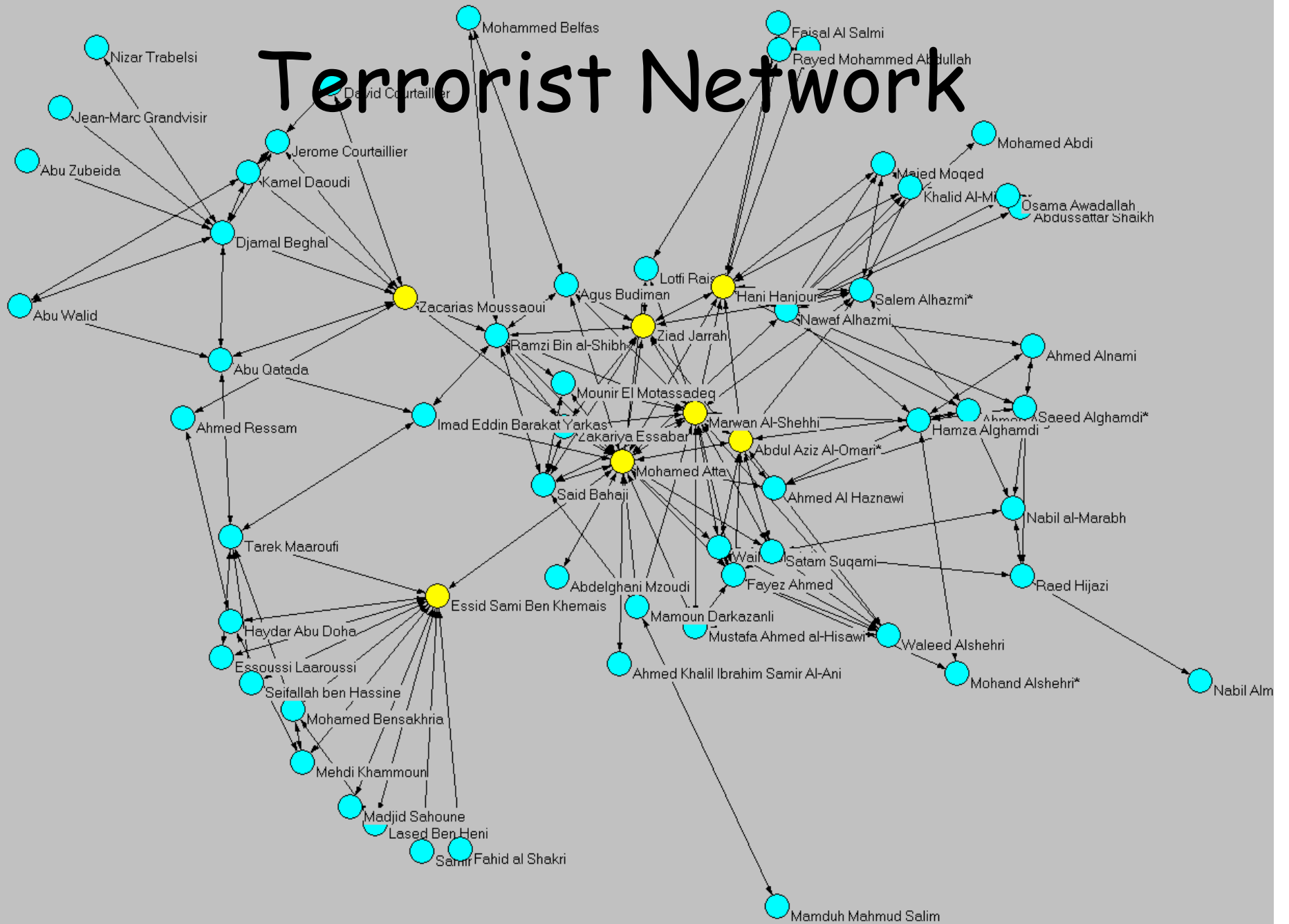
Outline

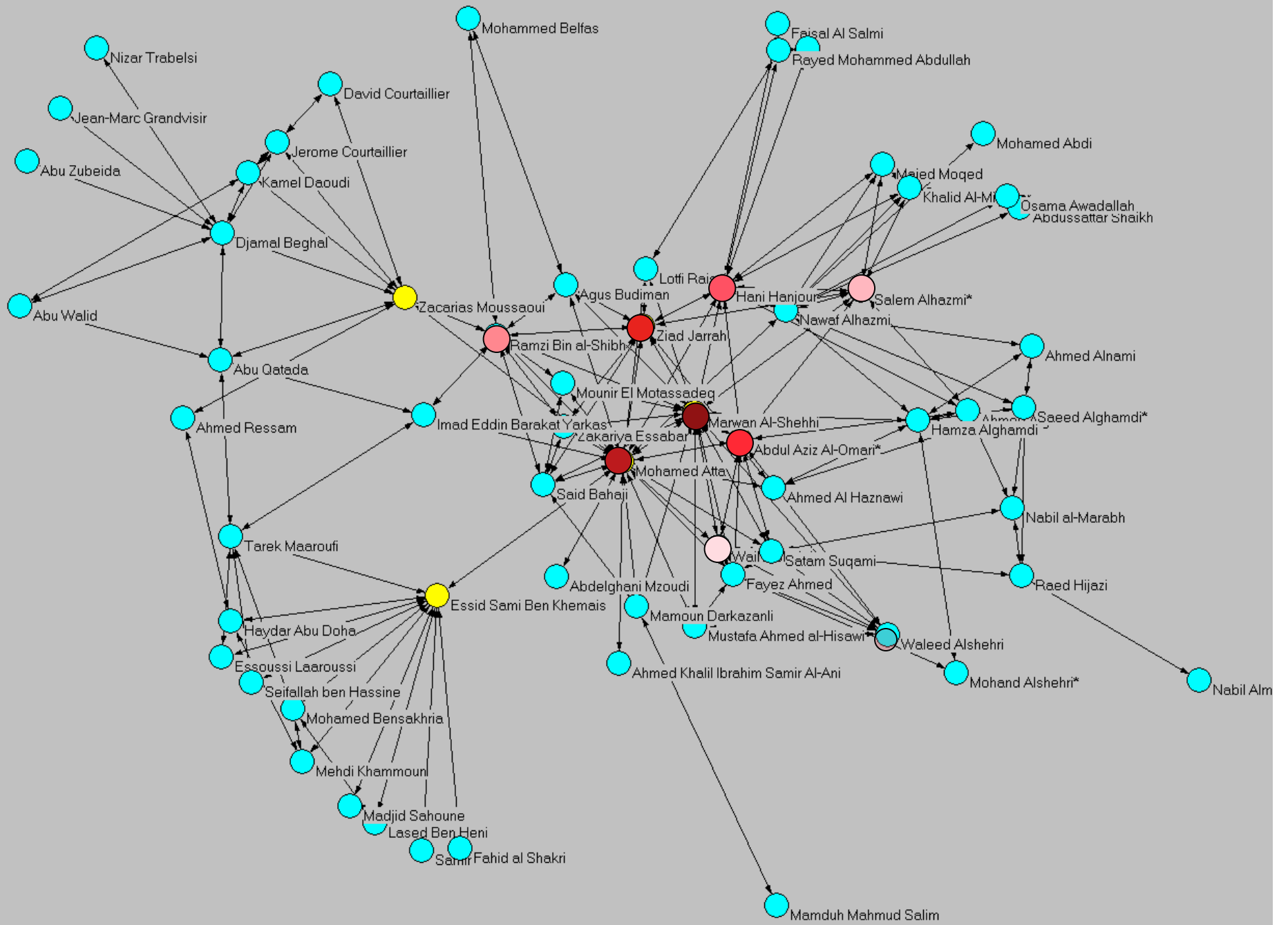
- Sudoku
- Military Applications
 - planning flight paths
 - disabling and herding communication in networks
- Ranking Applications
 - ranking on the World Wide Web
- Clustering and Data Mining Applications
 - clustering the Enron email dataset
 - clustering on terrorist networks

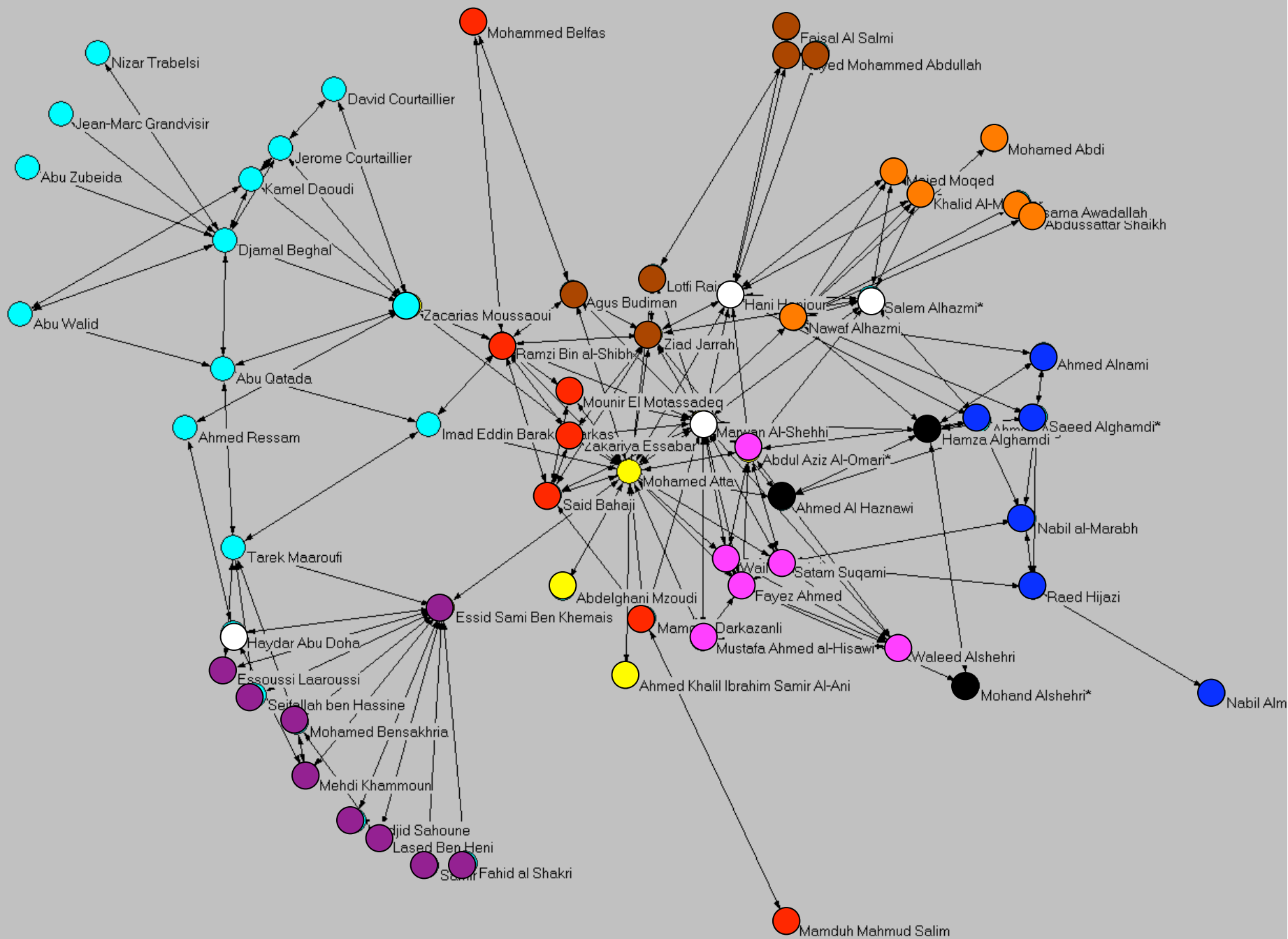
Data Mining on Terrorist Networks

- locating most important terrorists
- clustering terrorists
- identifying *central nodes* in terrorist network

Terrorist Network







Conclusions

- **Mathematics is useful.**

To isolate mathematics from the practical demands of the sciences is to invite the sterility of a cow shut away from the bulls. —P. L. Chebyshev

Mathematics is a more powerful instrument of knowledge than any other that has been bequeathed to us by human agency. —Descartes

- **Mathematical models scale well.**

2 radars vs. 200 radars: the mathematical model doesn't care.

- **Mathematical models are broadly applicable.**

Same mathematical techniques solve Sudoku, flight route, clustering problems.

There is no branch of mathematics, however abstract, which may not someday be applied to the phenomena of the real world. —N. Lobachevsky

- **Mathematical research is an inventive process, which takes time, and**

Conclusions

- **Mathematics is useful.**

To isolate mathematics from the practical demands of the sciences is to invite the sterility of a cow shut away from the bulls. —P. L. Chebyshev

Mathematics is a more powerful instrument of knowledge than any other that has been bequeathed to us by human agency. —Descartes

- **Mathematical models scale well.**

2 radars vs. 200 radars: the mathematical model doesn't care.

- **Mathematical models are broadly applicable.**

Same mathematical techniques solve Sudoku, flight route, clustering problems.

There is no branch of mathematics, however abstract, which may not someday be applied to the phenomena of the real world. —N. Lobachevsky

- **Mathematical research is an inventive process, which takes time, and**

$$Time = Money$$