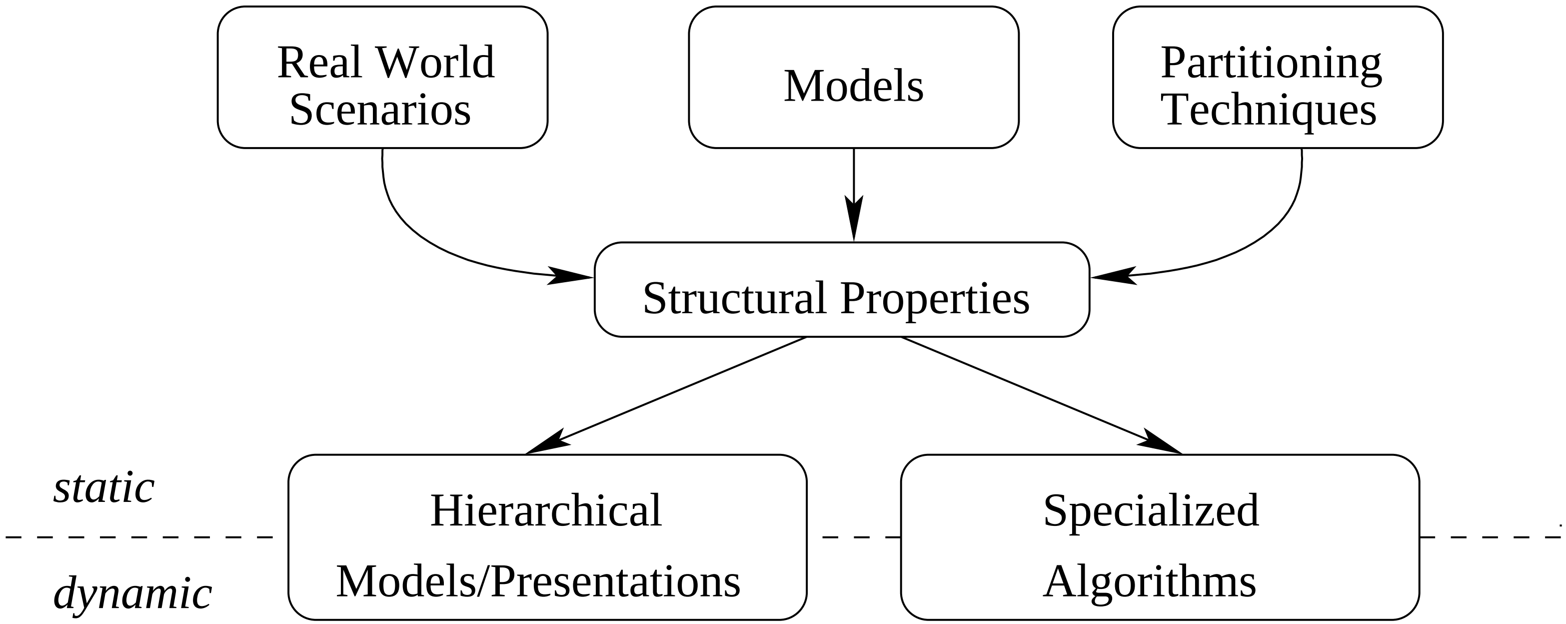
Density based clustering in dynamic and abstract representations of large networks

Klaus Holzapfel

Lehrstuhl für Effiziente Algorithmen
Fakultät für Informatik an der Technische Universität München

Jahrestreffen 22. – 24. Juli 2002 (Konstanz)

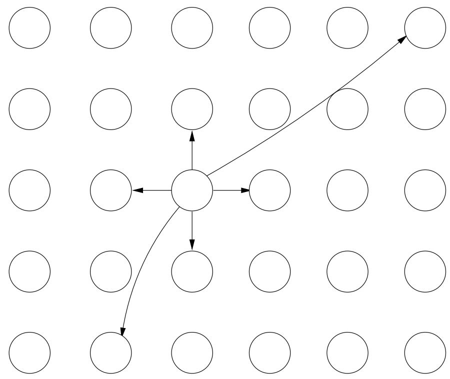
Project Overview



Semi-structured Data – Models & Algorithms

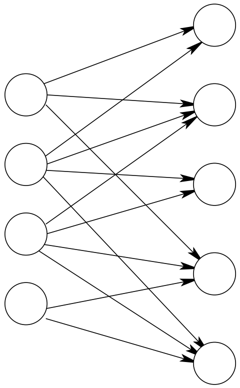
- Kleinberg [STOC 00]

  Transportation Problem
  (only local information)
  Algorithm: $\mathcal{O}(log^2 n)$



- Kleinberg [J. ACM 99]

  Hubs and Authorities



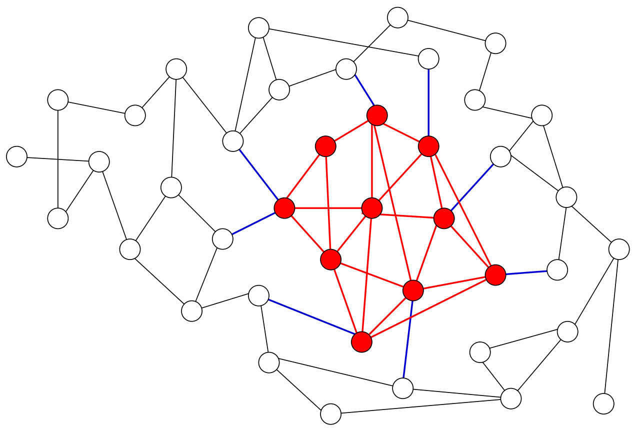
- Achlioptas, Fiat, Karlin, McSherry [FOCS 01]

  Web Search via Hub Synthesis

Practical Usage of Abstracting Data

- VLSI Design
  - placement
  - routing and wiring
- Transportation Problems
  - telephone network
  - road network
- Clustering
  - 3-D data representation (simulators)
  - speech recognition
  - web communities

How to cluster data?



- many internal edges (density)
- few external edges (cut)
- different short paths (connectivity)

Problem:	DENSE k-SUBGRAPH-PROBLEM
Input:	Graph G, $k \in \mathbb{N}$
Output:	Subgraph G' having maximum number of edges w.r.t. all subgraphs of size k

- (variable) decision problem $\mathcal{NP}$-complete
- $\mathcal{O}(n^{\frac{1}{3}-\epsilon})$-approximation [Feige, Kortsarz, Peleg, 2001]

γ-Dense Subgraph Problem

Problem:	γ-DENSE SUBGRAPH-PROBLEM (γ-DSP)
Input:	Graph G, $k \in \mathbb{N}$
Output:	Does there exist a subgraph G' of size k having at least $\gamma(k)$ edges

- $\gamma(k) = \binom{k}{2}$ $\quad$ γ-DSP $=$ CLIQUE $\in \mathcal{NP}$-c
- $\gamma(k) = 0$ $\quad$ γ-DSP $\in \mathcal{P}$

Where is the threshold?

Results – Overview

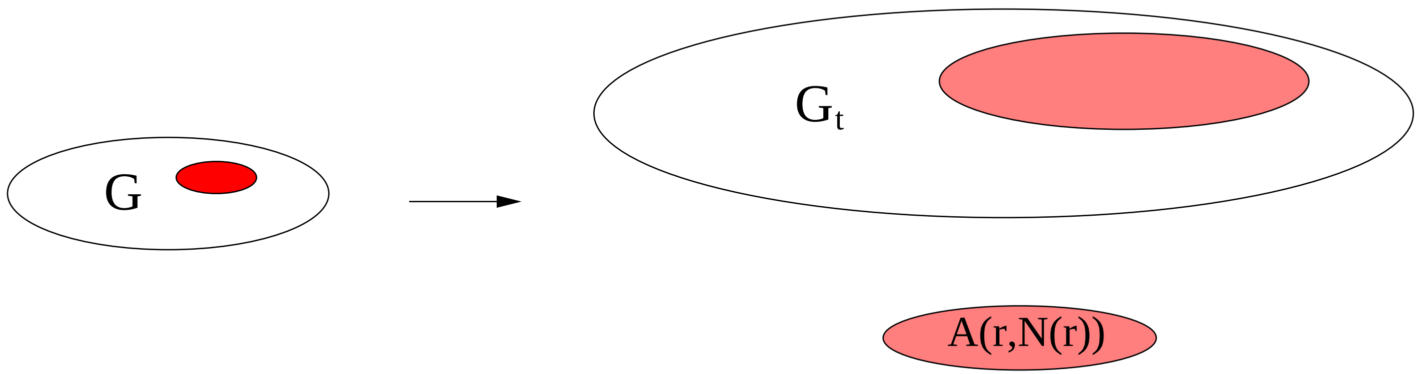
	$\mathcal{P}$	$\mathcal{NP}$-c
[Asahiro et.al. 2002]	$\gamma(k) = k$	$\gamma(k) = \Theta(k^{1+\epsilon})$
[Feige, Seltser 1997]		$\gamma(k) = k + k^{\epsilon}$
[H et.al. 2002]	$\gamma(k) = k + O(1)$	$\gamma(k) = k + \Theta(k^{\epsilon})$

$\mathcal{NP}$-completeness

$\mathcal{NP}$-c Theorem

Theorem. *The* γ-DSP *is* $\mathcal{NP}$*-complete for* $\gamma(k) = k + \Theta(k^\epsilon)$ *(*γ *must be be computable in polynomial time;* $0 < \epsilon < 2$*).*

Proof sketch: • CLIQUE$_{\frac{1}{2}} \preceq_m^p \gamma$-DSP



$$
\begin{aligned}
N(r) &= \gamma\left(k + t\binom{k}{2} + r\right) - (t+1)\binom{k}{2} \\
r &= 30D^2k \\
t &= \left\lceil (6D)^{3\epsilon^{-1}} k^{2(1-\epsilon)\epsilon^{-1}} \right\rceil
\end{aligned}
$$

Polynomial Time Algorithm

Polynomial Time Algorithm

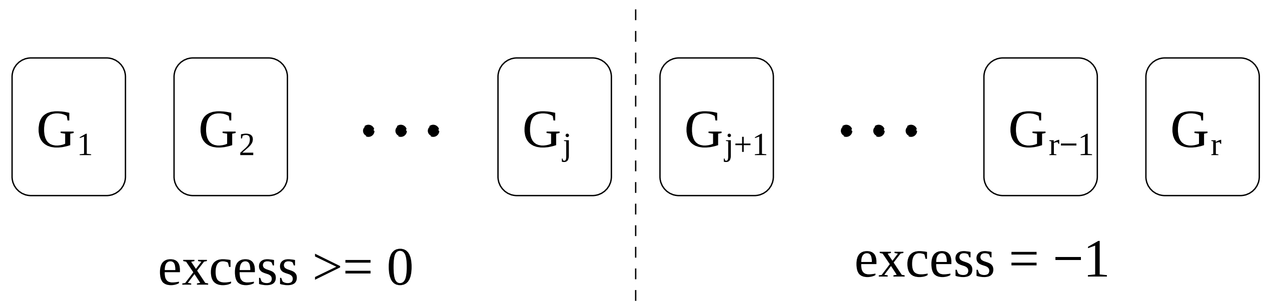
consider: $\gamma(k) = k + \mathcal{O}(1)$
definition: $\mathrm{excess}(G) = |E(G)| - |V(G)|$

Problem:	EXCESS-c-SUBGRAPH
Input:	Graph G, $k \in \mathbb{N}$
Output:	Does G contain a subgraph G' of size k and $\mathrm{excess}(G') = c$?

Theorem. *Given G and $k \in \mathbb{N}$, the problem* EXCESS-c-SUBGRAPH *can be solved in time $\mathcal{O}(|V|^{2c+3})$.*

Proof — Sketch

- Sort the connected components by their excess (decr.)



(1) negative components are necessary

(2) positive components are sufficient

$\rightarrow$ compute $A_i[..]$ and use dynamic programming

# of vertices	1	2	3	$\ldots$	5	$\ldots$	$\lvert V(G_i)\rvert$
max. excess	-1	-1	0	$\ldots$	2	$\ldots$	$\star$

$\star = \min(\text{excess}(G_i), c+1)$

Proof — How to find $A_i[x]$

Consider a vertex minimal subgraph $G_{\min}$ with excess c

- $$\sum_{v \in V(G_{\min})} \deg_{G_{\min}}(v) = 2\|E(G_{\min})\| = 2(\|V(G_{\min})\| + c)$$
- In $G_{\min}$ there is no vertex with degree less than 2, thus:

  $$\sum_{v \in V(G_{\min})} (\deg_{G_{\min}}(v) - 2) = 2(\|V(G_{\min})\| + c) - 2\|V(G_{\min})\| = 2c$$

  Therefore, the number of vertices with degree ≥ 3 is at most $2c$, i.e. $\mathcal{O}(n^{2c})$ possible combinations.

  $\Rightarrow$ enumeration possible in polynomial time
- Each such combination can be tested using parallel BFS.

Calculation of A_i can be done in time $\mathcal{O}(n^{2c+3})$.

Theorem. *Let* $\gamma : \mathbb{N} \rightarrow \mathbb{N}$ *be a function that is computable in polynomial time:*

1. *If* $\gamma(k) = k + \mathcal{O}(1)$ *then* γ-DSP *is in* $\mathcal{P}$.
2. *If* $\gamma(k) = k + \Theta(k^{\epsilon})$*, for some rational number* $0 < \epsilon < 2$*, then* γ-DSP *is* $\mathcal{NP}$*-complete.*

Directed Graphs

How to measure density in directed graphs ?

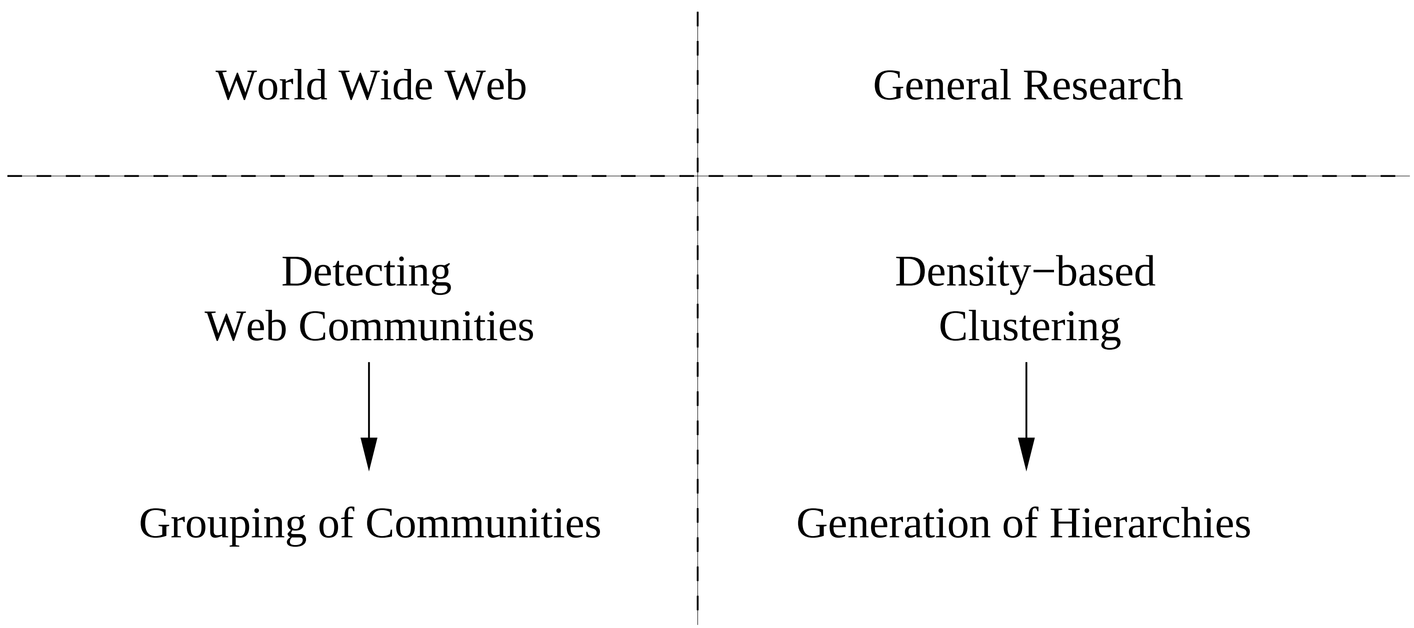
- directed graphs [Kannan, Vinay, 1999]:

$$\delta(G) = \frac{2|E(G)|}{|V(G)|} \quad \Rightarrow \quad \delta'(G) = \max_{S,T \subseteq V(G)} \left(\frac{E(S,T)}{\sqrt{|S||T|}} \right)$$

- evaluating existence of good hubs and authorities
- S and T not disjoint
- how to proceed when searching for dense bipartite graphs?

Future Work (1)

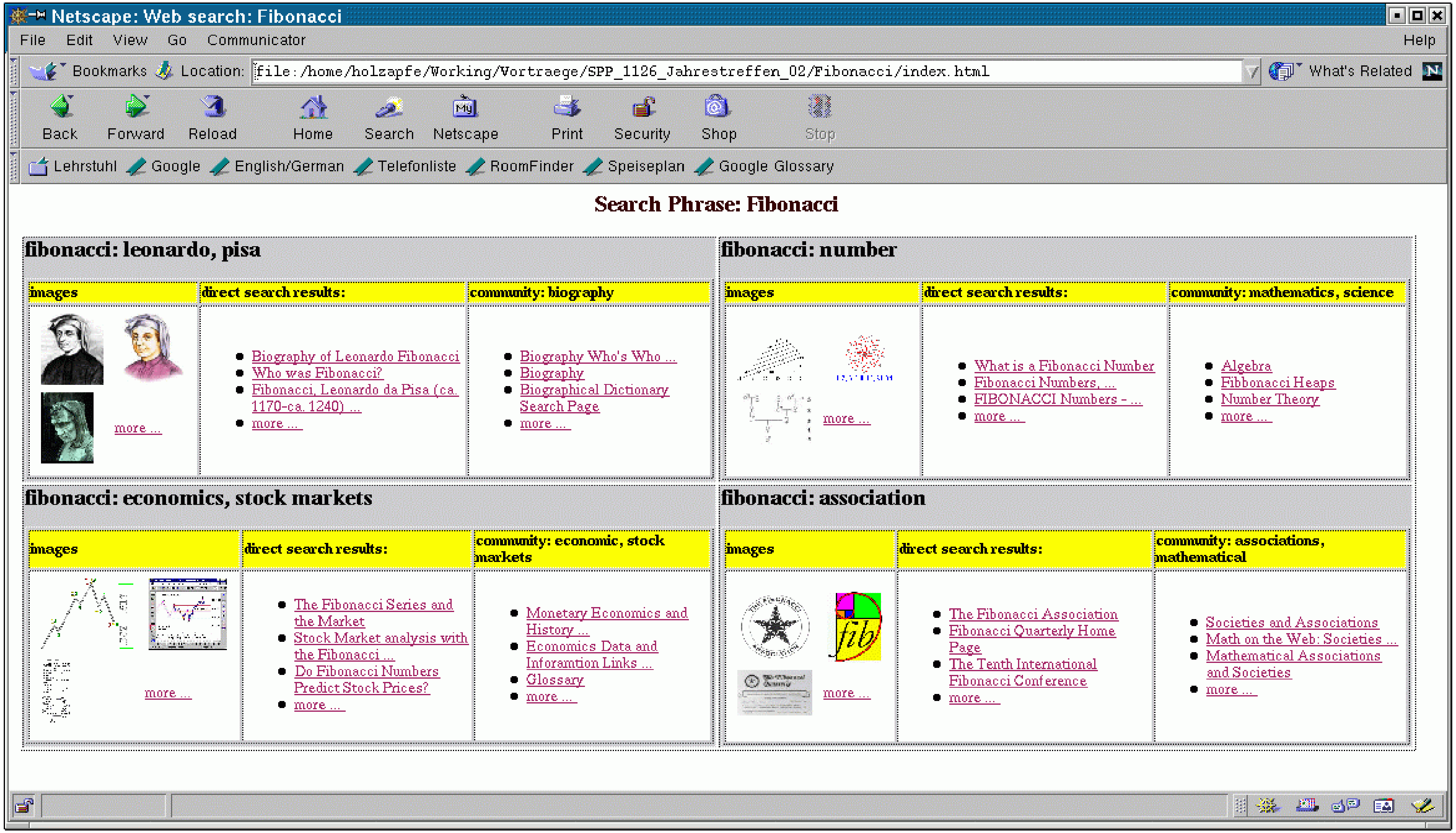
Where to go? What to do next?



- *Trawling the Web for Emerging Cyber Communities*
  [Kumar, Raghavan, Rajagopalan, Tomkins, 1999] WWW8
- *An approach to build a cyber-community hierarchy*
  [Krishan Reddy, Kitsuregawa, 2002] Workshop on Web Analytics

How to utilize a Hierarchy of Web Communities

Web search for: Fibonacci



Theoretical Aspects

Algorithmic — How good can problems be approximated within different types of hierarchies and graph classes?

- shortest paths, local vs. global
- distance and connectivity
- searching and similarity

Dynamic aspects in hierarchies — Real world systems are not static; objects and relationships vary over time.

- recognition of emerging / dissolving clusters
- re-calibration of cluster properties (weight, size, ...)
- local vs. global recalculation