

Visualizing Data with Graphs and Maps

Yifan Hu

AT&T Labs – Research

NIST
May 7, 2012

Outline

- The graph visualization problem
- Algorithms & challenges for visualizing large graphs
- Visualizing cluster relationships as maps

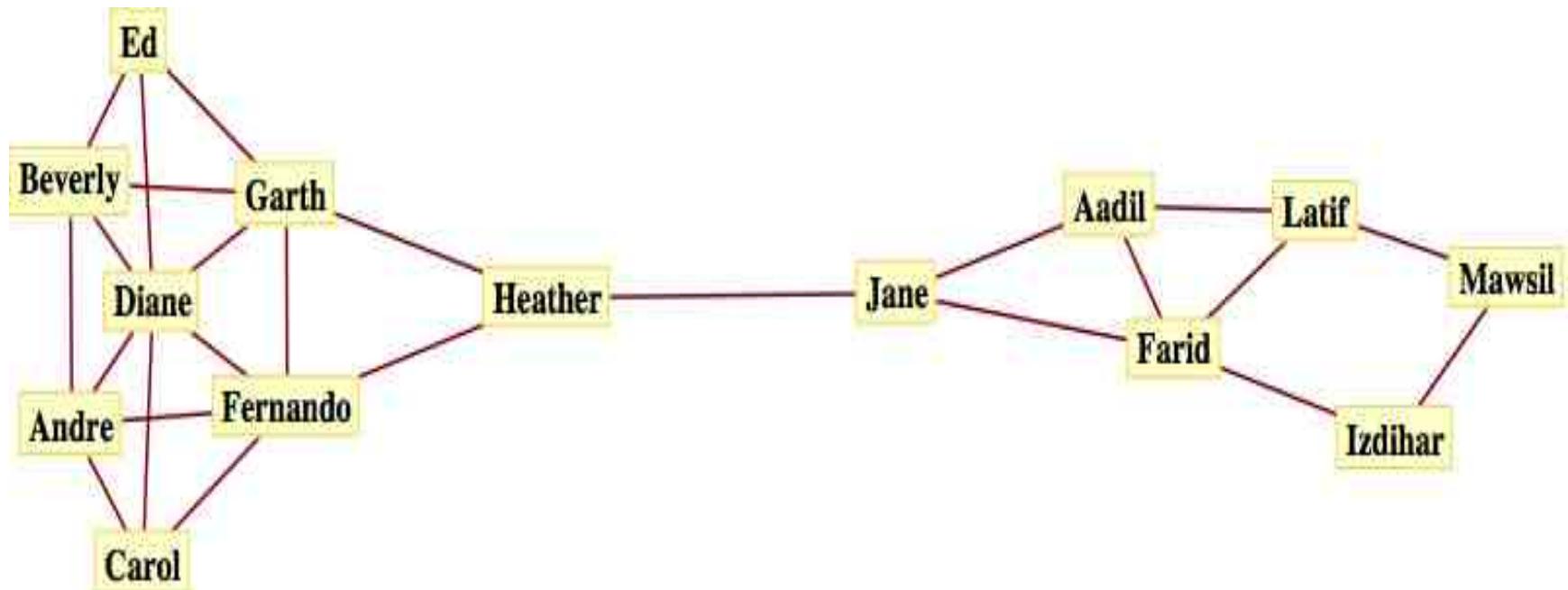
The graph visualization problem

- Given some relational data

{Farid—Aadil, Latif—Aadil, Farid—Latif,
Carol—Andre, Carol—Fernando, Carol—Diane, Andre
—Diane, Farid—Izdihar, Andre—Fernando, Izdihar—
Mawsil, Andre—Beverly, Jane—Farid, Fernando—
Diane, Fernando—Garth, Fernando—Heather, Diane—
Beverly, Diane—Garth, Diane—Ed, Beverly—Garth,
Beverly—Ed, Garth—Ed, Garth—Heather, Jane—Aadil,
Heather—Jane, Mawsil—Latif}
- It is not easy to see what's going on!

The graph visualization problem

- But if we visualize it

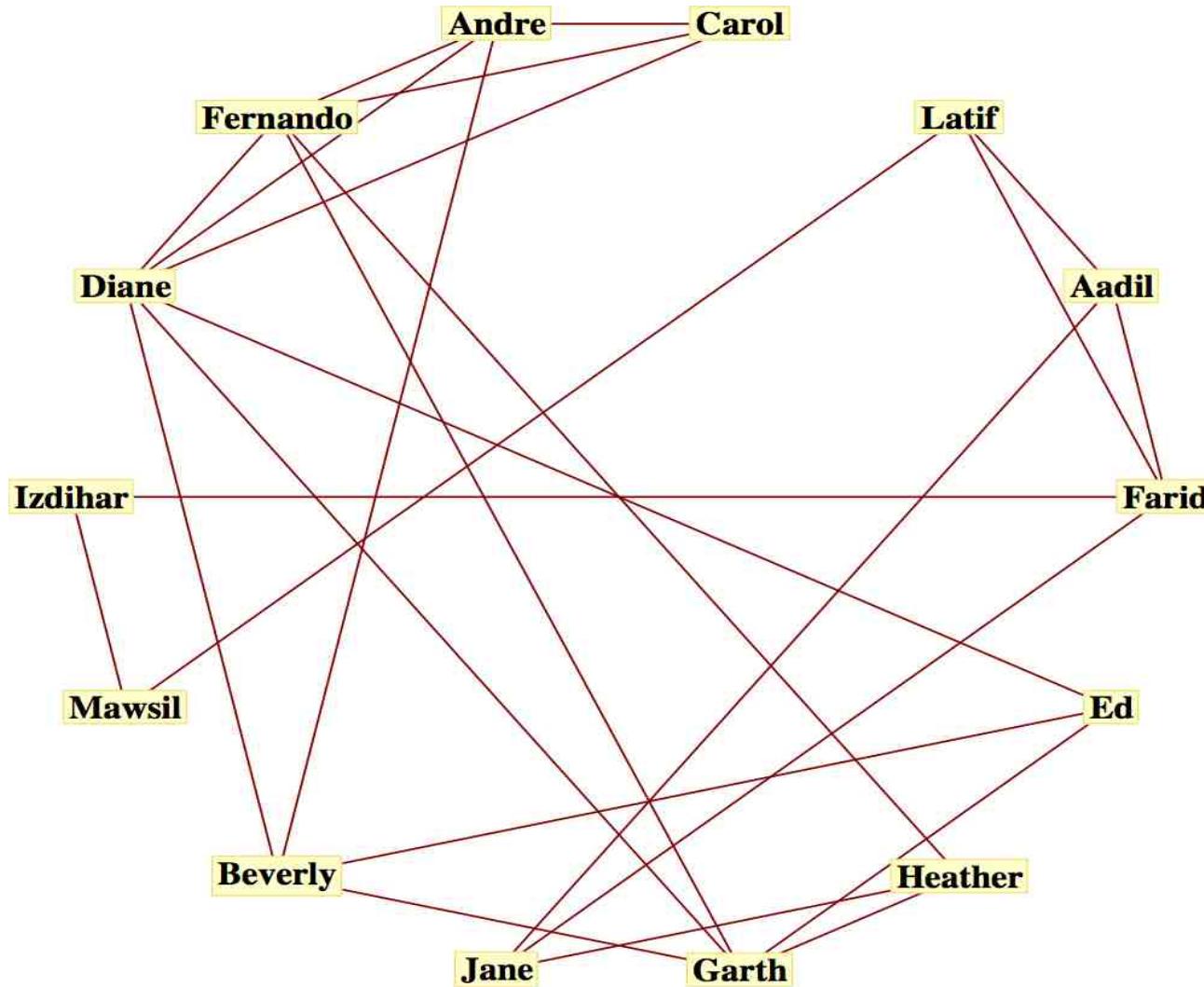


The graph visualization problem

- The graph visualization problem: to achieve a “good” visual representation of a graph using node-link diagram (points and lines).
- Main criteria for a good visualization: readability and aesthetics.
- Small area, good aspect ratio, few edge cross-overs, showing symmetry/clusters if exist, sufficiently large edge-edge, node-node and node-edge resolution, planar drawing for planar graph, ...

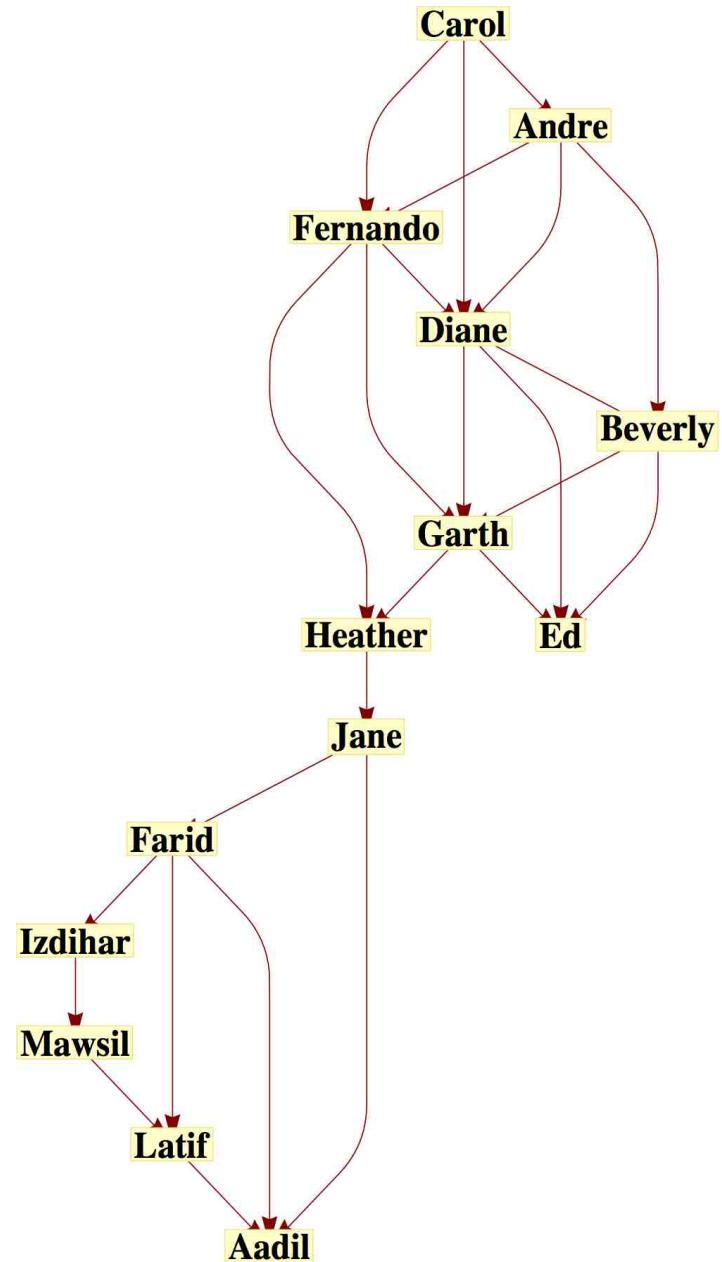
The graph visualization problem

- Different styles of graph drawing: circular layout



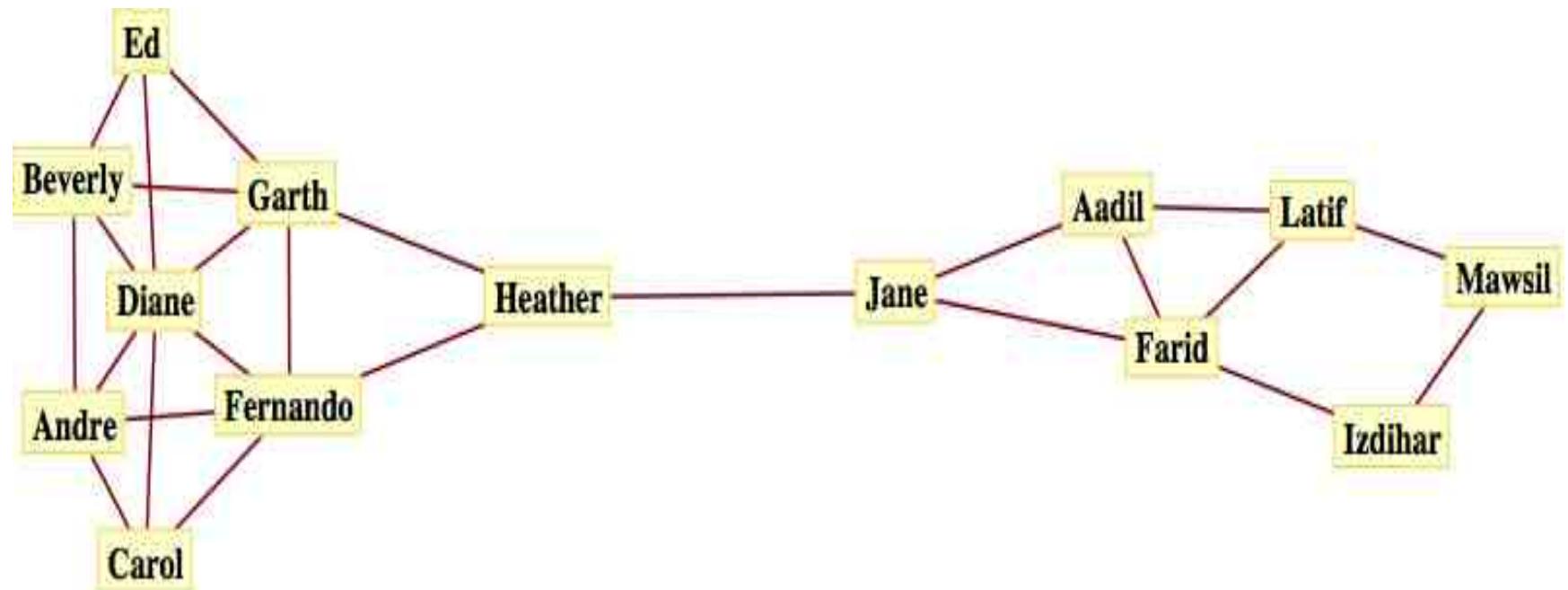
The graph visualization problem

- Different styles of graph drawing: hierarchical layout



The graph visualization problem

- Other styles: orthogonal, grid drawing, visibility drawings.
- This talk concentrates on undirected/straight edge drawing of non-planar graphs.

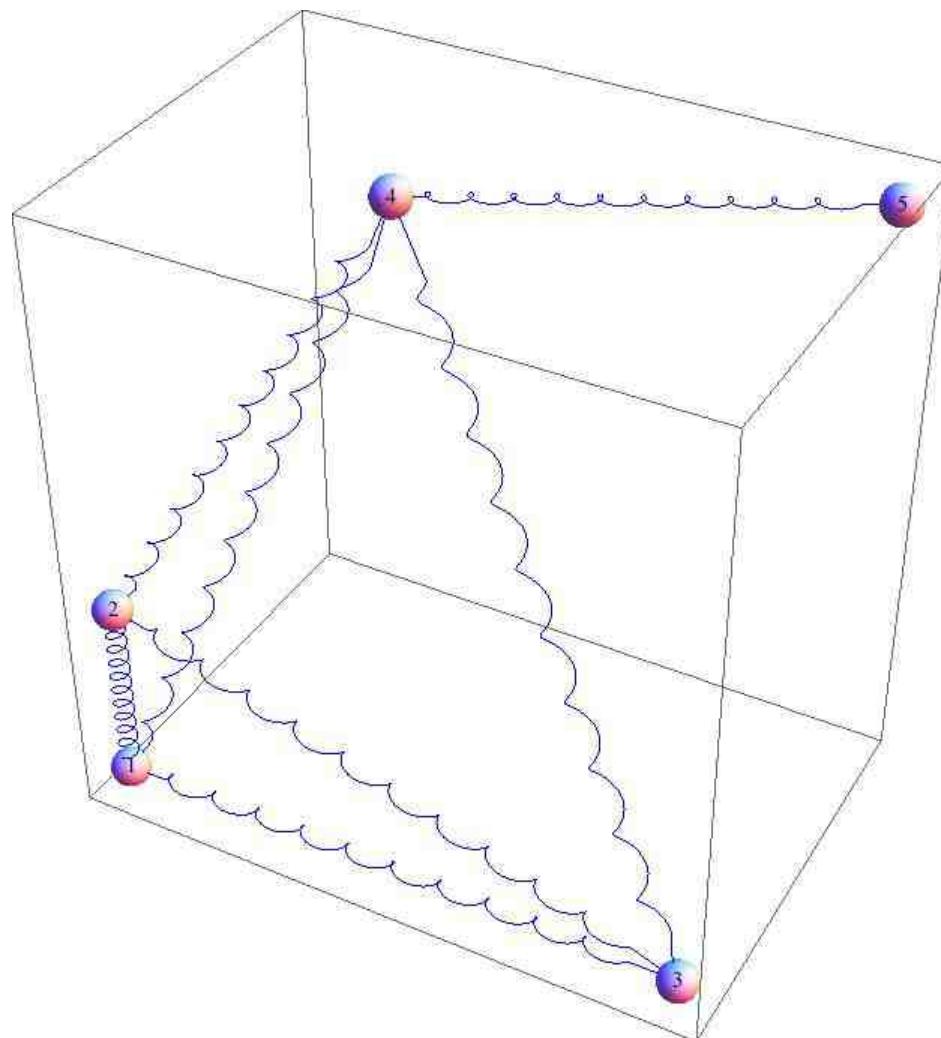


Graph drawing algorithms

- Hand layout not feasible (unless small graphs)
- Automated algorithms needed
- Virtual physical models are popular
- Spring model vs spring-electrical model
- Spring model: a spring between every pair of vertices
- Ideal spring length = graph distance

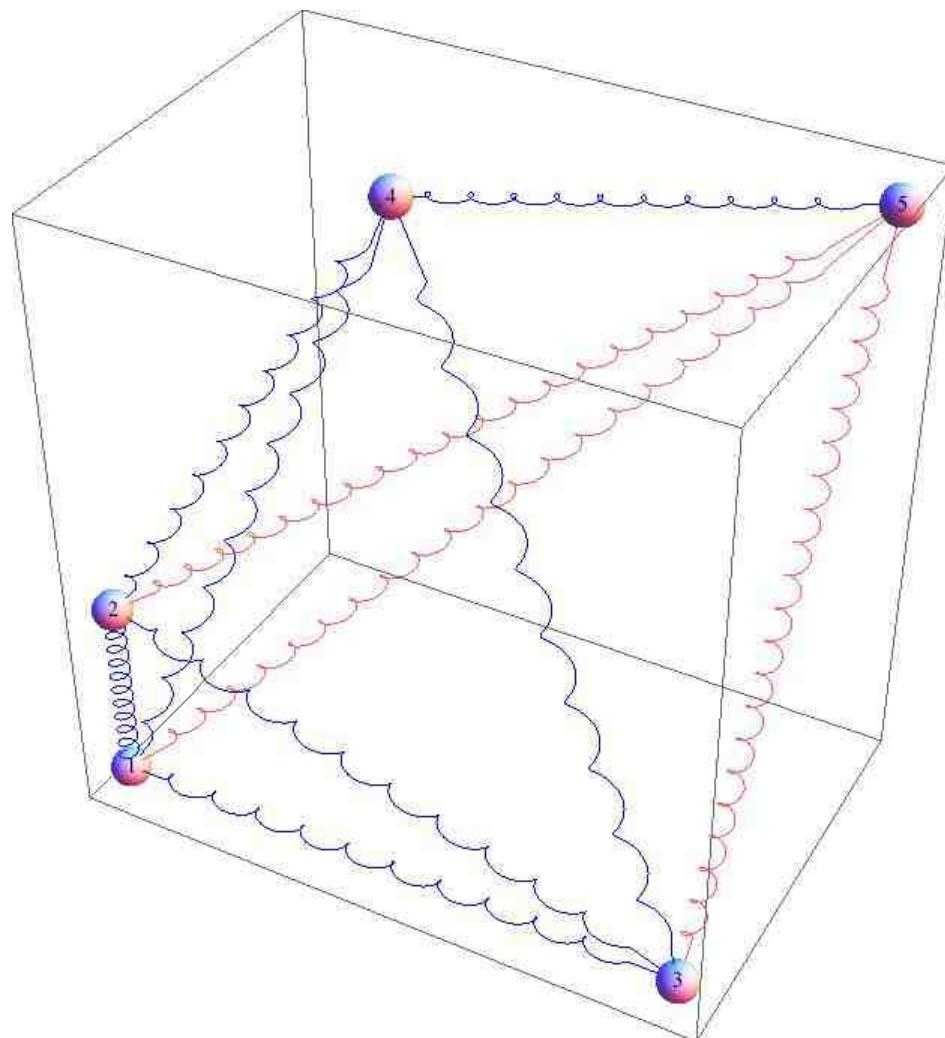
Spring Model (aka Stress Model)

- {1—2, 2—3, 1—3, 1—4, 2—4, 3—4, 4—5}



Spring Model (aka Stress Model)

- {1—2, 2—3, 1—3, 1—4, 2—4, 3—4, 4—5}

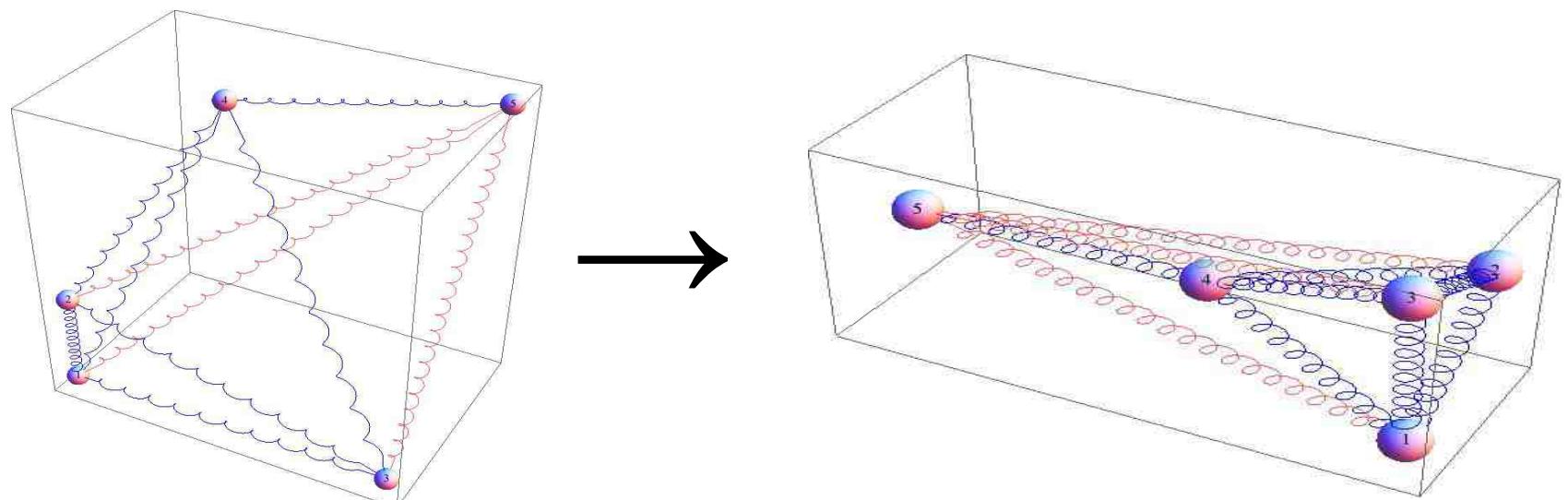


Spring Model (aka Stress Model)

- Spring model

$$\text{stress}(x) = \sum_{i \neq j, i, j \in V} w_{ij} (\| x_i - x_j \| - d_{ij})^2$$

- Kruskal & Seery (1980); Kamada & Kwai (1989)



Spring Model (aka Stress Model)

- Spring model

$$\text{stress}(x) = \sum_{i \neq j, i, j \in V} w_{ij} (\| x_i - x_j \| - d_{ij})^2$$

- Solution method:
- $$x_i \leftarrow \frac{\sum_{j \neq i} w_{ij} \left(x_j + d_{ij} \frac{x_i - x_j}{\|x_i - x_j\|} \right)}{\sum_{j \neq i} w_{ij}}$$
- $$L_w x := L_d x$$

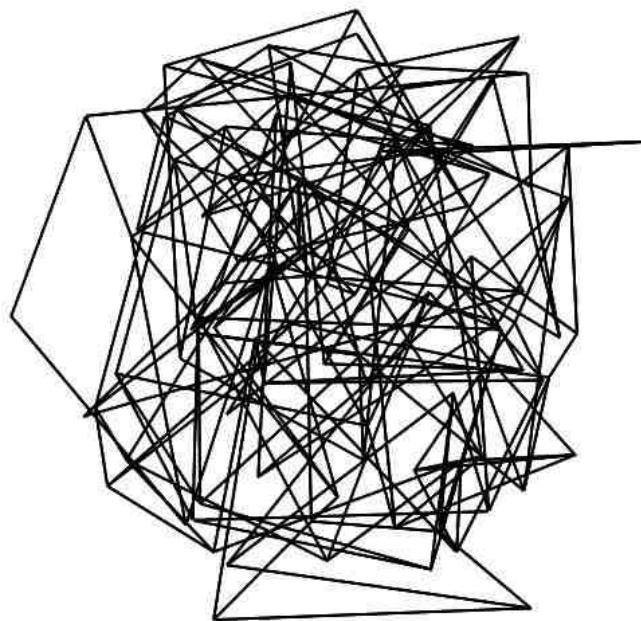
L_w : weighted (dense) Laplacian

$$(L_d x)_i = \sum_{j \neq i} w_{ij} d_{ij} \frac{x_i - x_j}{\|x_i - x_j\|}$$

- Stress majorization (de Leeuw, J. , 1977; Gasner, Koren & North, 2004)

Spring Model (aka Stress Model)

- Stress majorization on a grid graph



pmds(k), iter=1. 100 nodes, 360 edges.

Spring Model (aka Stress Model)

- Stress majorization on a grid graph



Spring Model (aka Stress Model)

- But this model is not scalable
- All-pairs shortest paths: $O(|V|^2 \log |V| + |V||E|)$
- Memory: $O(|V|^2)$

Spring-electrical Model

- Eades (1984), Fruchterman & Reigold (1991)
- Energy to minimize:

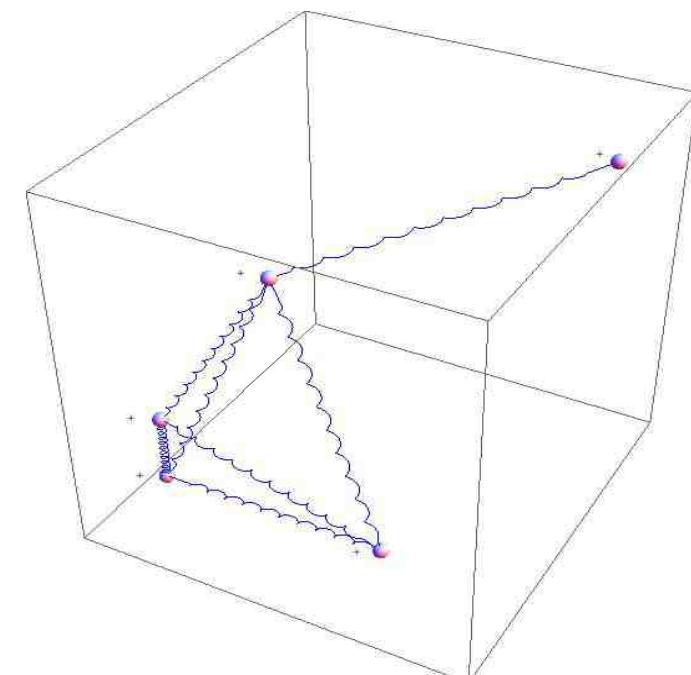
$$\frac{\sum_{i \leftrightarrow j} \|x_i - x_j\|^3}{3K} - K^2 \sum_{i \neq j} \ln (\|x_i - x_j\|)$$

- Repulsive force =

$$-\frac{K^2}{\|x_i - x_j\|} \frac{x_i - x_j}{\|x_i - x_j\|}, \quad i \neq j$$

- Attractive force =

$$\frac{\|x_i - x_j\|^2}{K} \frac{x_i - x_j}{\|x_i - x_j\|}, \quad i \leftrightarrow j$$

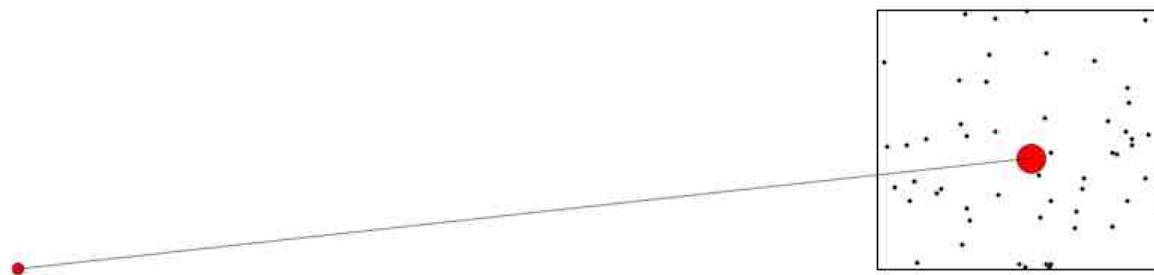


Spring-electrical Model

- Force directed iterative process:
 - for every node
 - calculate the attractive & repulsive forces
 - move the node along the direction of the force
 - repeat until converge
- But still not scalable: all-to-all repulsive force
 - $$-\frac{K^2}{\|x_i - x_j\|} \frac{x_i - x_j}{\|x_i - x_j\|}, \quad i \neq j \quad O(|V|^2)$$
- Easy to get trapped in a local minima

Reducing the $O(|V|^2)$ complexity

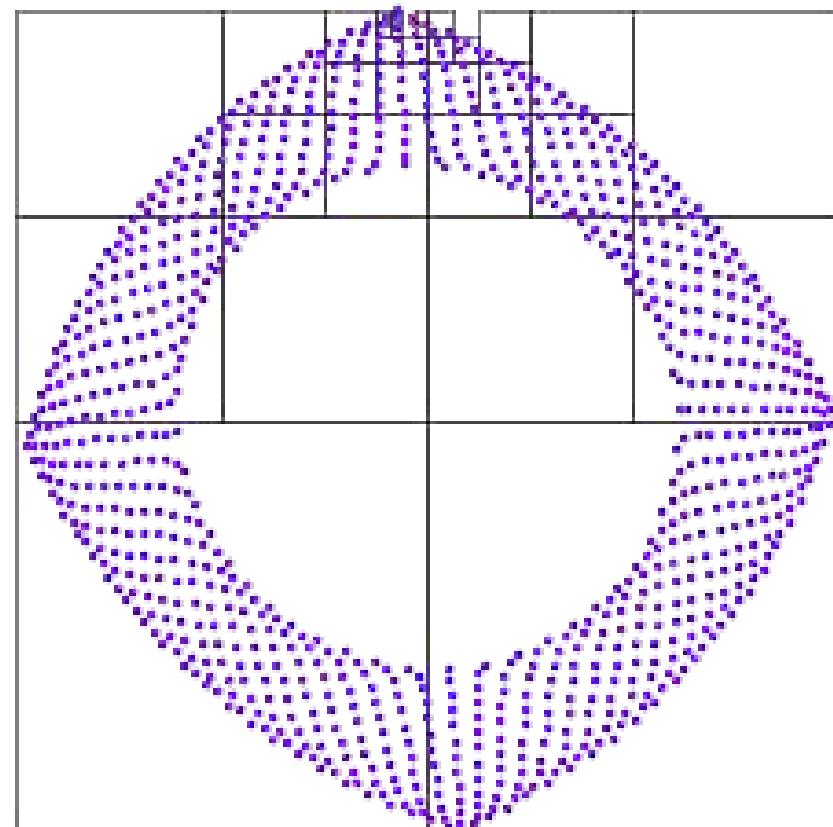
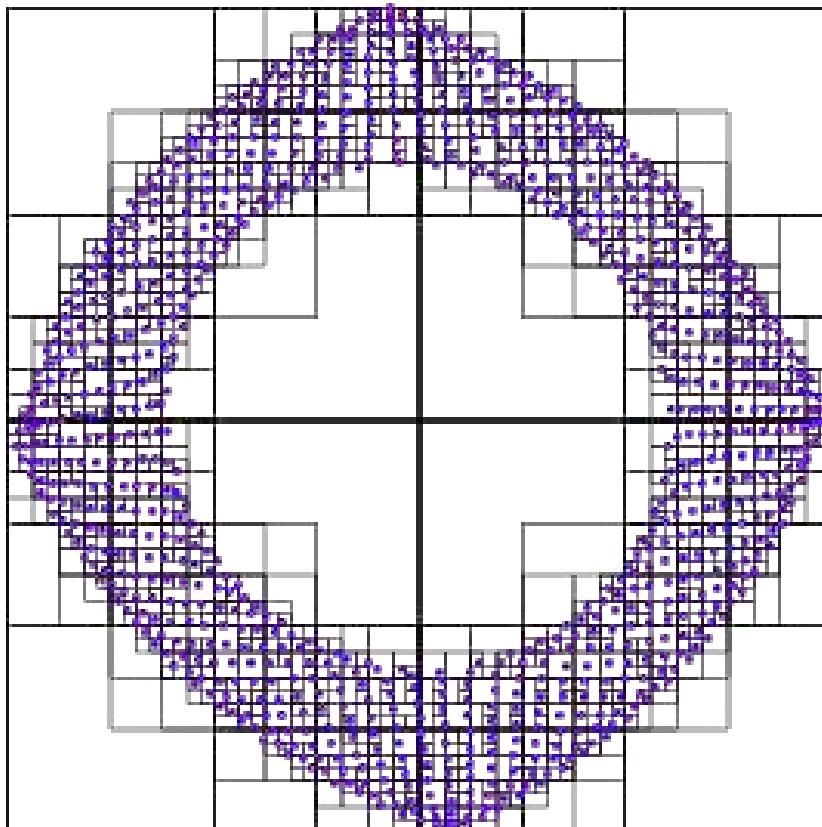
- Group remote nodes as supernodes
(Barnes-Hut, 1986; Tunkelang, 1999; Quigley 2001)



- Reduce complexity to $O(|V| \log(|V|))$

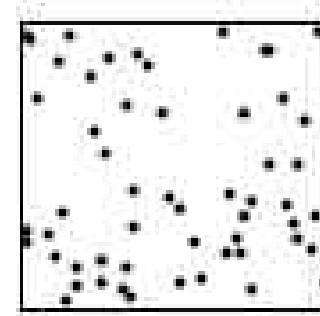
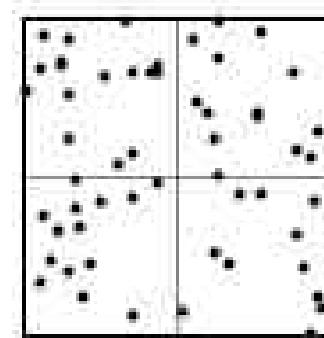
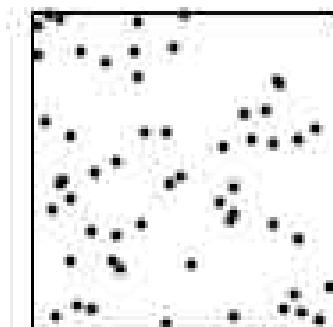
Reducing the $O(|V|^2)$ complexity

- Implementation: quadtree/KD-tree.
- Example: 932 → 20 force calculation.



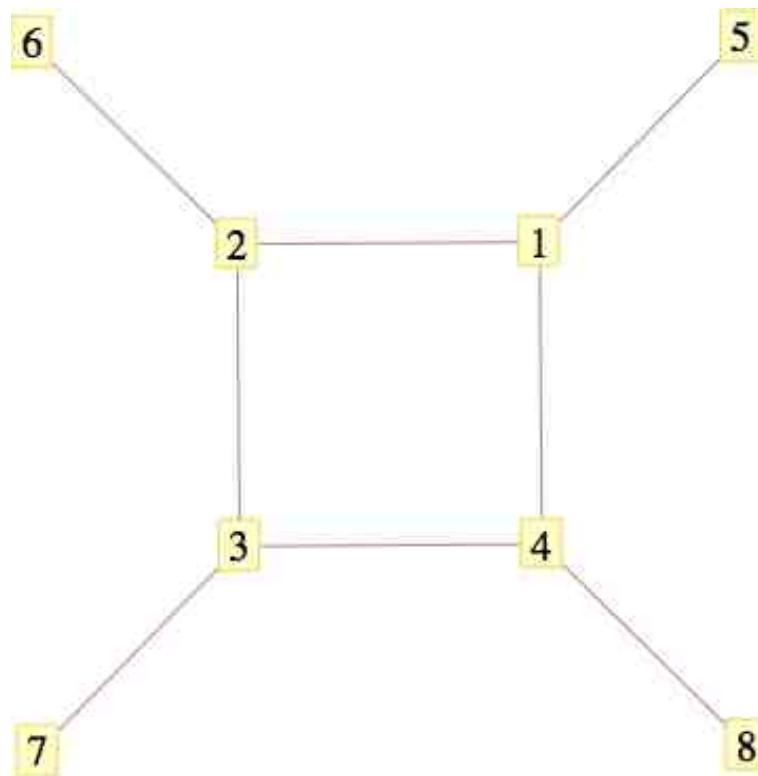
Reducing the $O(|V|^2)$ complexity

- Taking one step further: supernode-supernode.
- Burton et al. (1998), particle simulation.



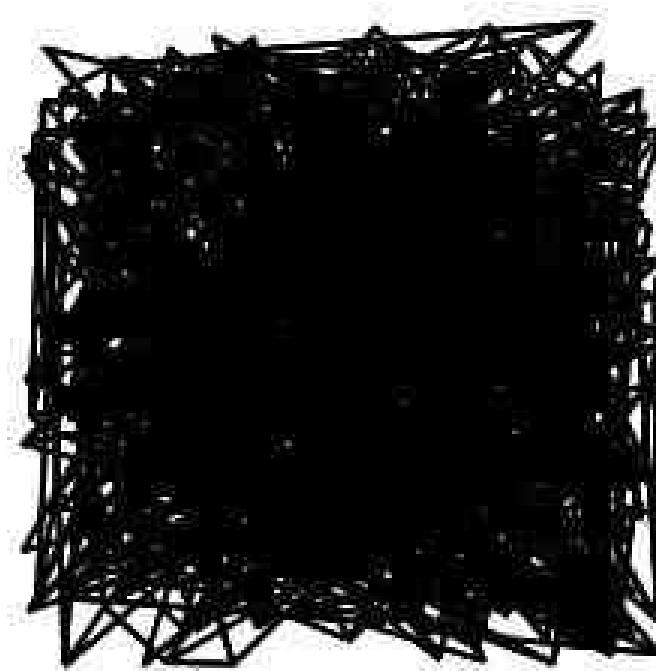
Finding global optimum

- Force directed algorithm: easy to get trapped in local min

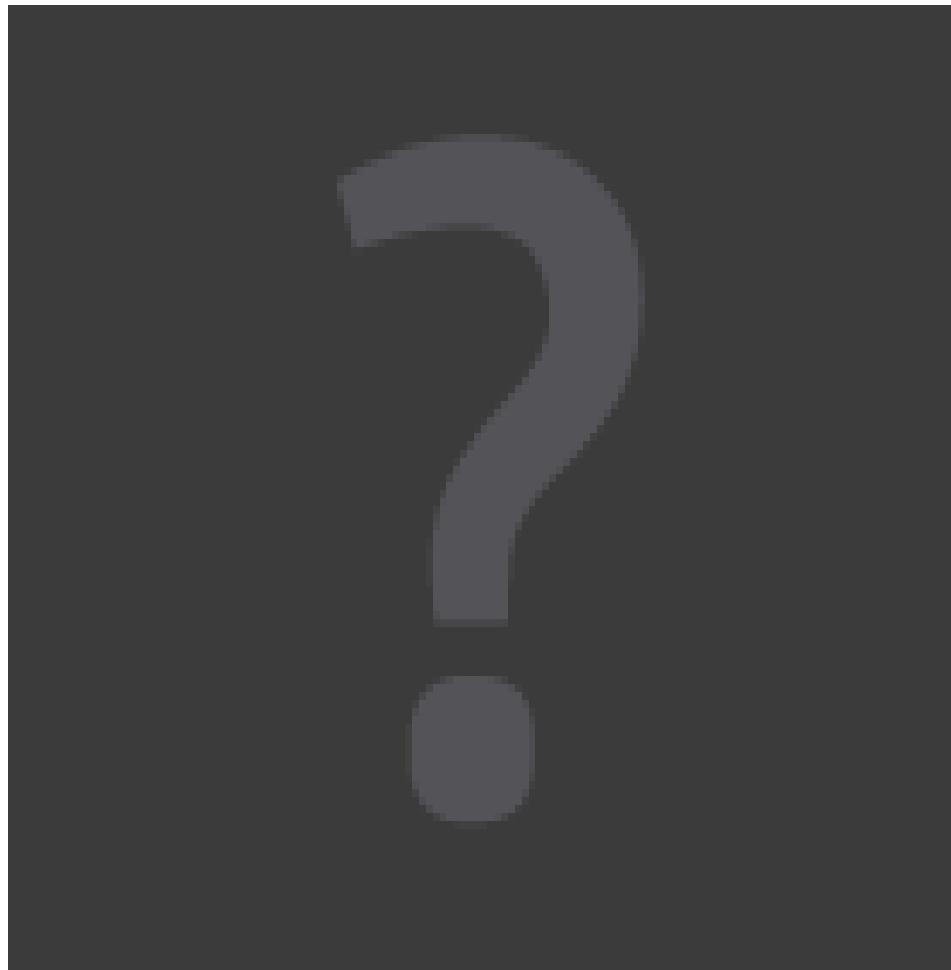


- The larger the graph, the more likely to get trapped.
- Also, smooth errors are harder to erase with iterative scheme

Finding global optimum

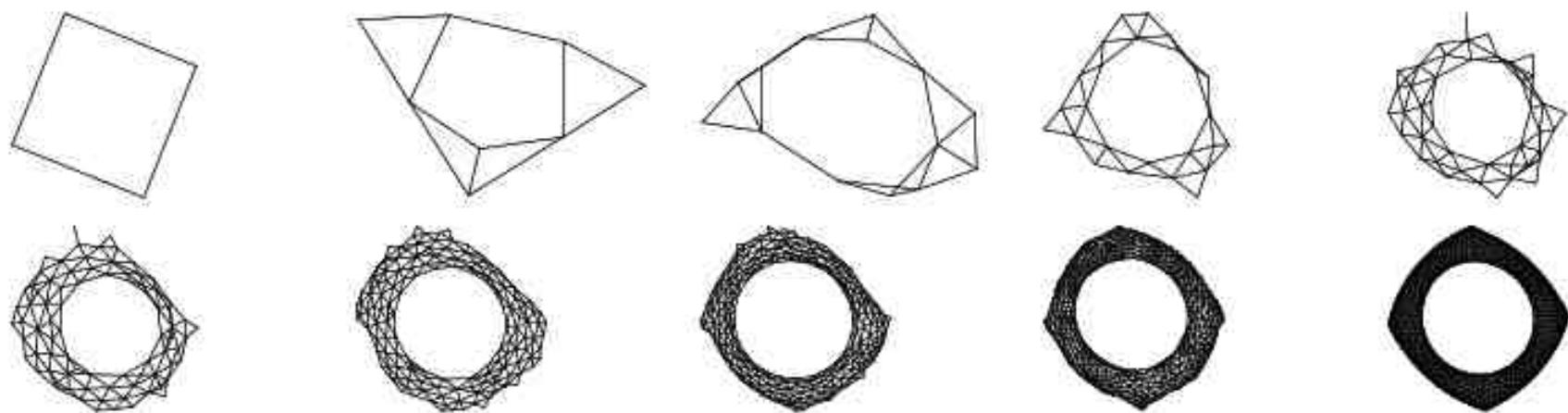


Finding global optimum



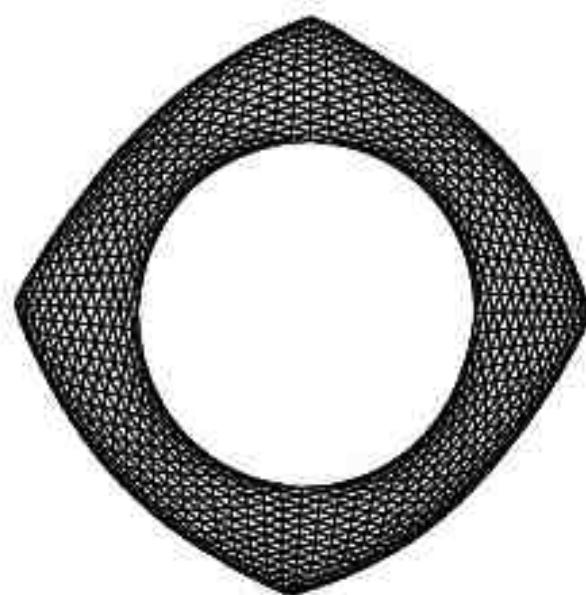
Global Optimum: Multilevel

- Global optimum more likely with multilevel approach (Walshaw, 2005)



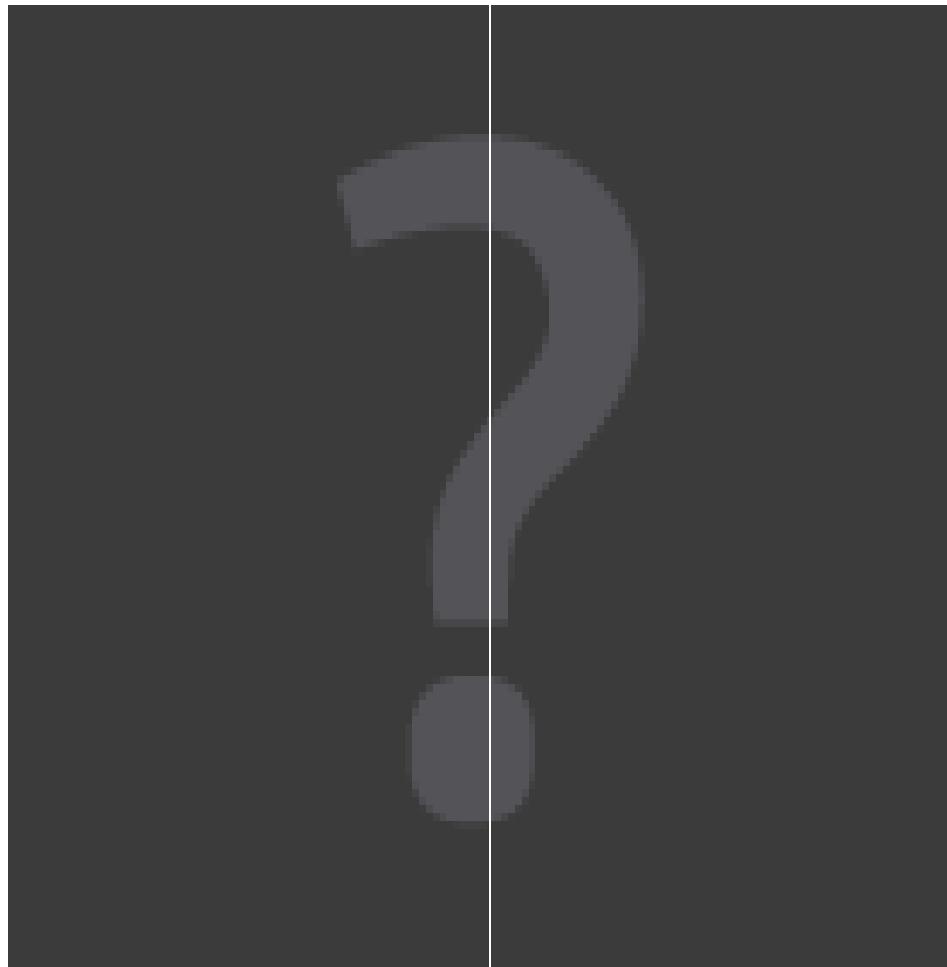
Spring-electrical: Large Graphs

- Multilevel + fast $O(|V|\log(|V|))$ force approximation → efficient & good quality graph layout algorithms (Hachul&Junger 2005; Hu 2005).



Spring-electrical: Large Graphs

- Multilevel + fast $O(|V|\log(|V|))$ force approximation → efficient & good quality graph layout algorithm (Hachul&Junger 2005; Hu 2005).



Other graph layout algorithms

- Eigenvector based methods (Hall's algorithm).

$$\min \sum_{i \leftrightarrow j} \|x_i - x_j\|^2, \text{ subject to } \sum_{i \in V} \|x_i\|^2 = 1$$

$Lx = \lambda x$, $\lambda > 0$ and λ as small as possible

- High dimensional Embedding (Harel & Koren, 2002)

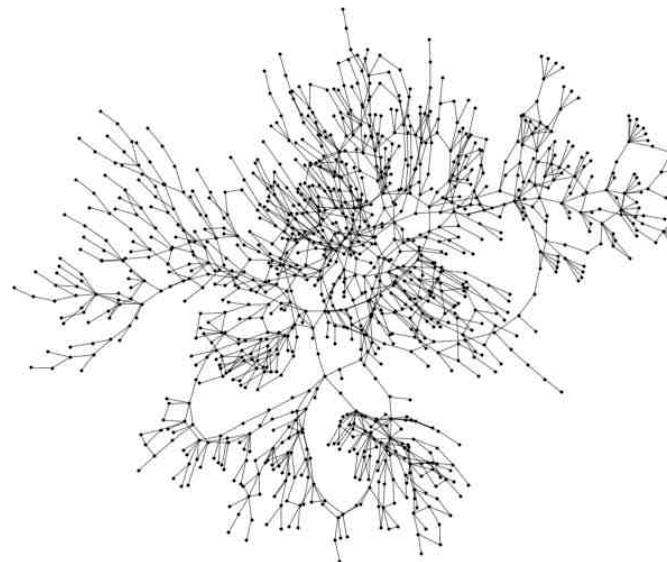
- Find distance from k vertices to all vertices

- Apply PCA to the $|V| \times k$ matrix to get the top 2 eigenvectors, use as coordinates

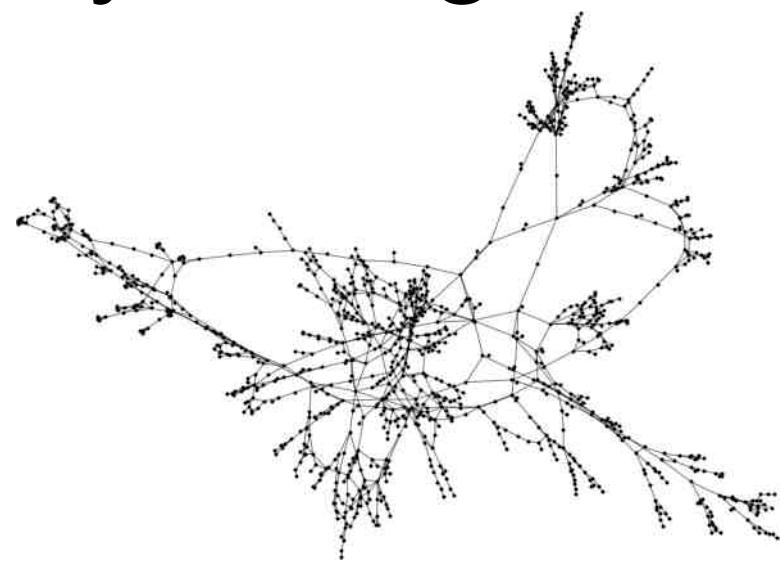
- PivotMDS (Brandes & Pich, 2006)

- All fast, but not good layout for graphs of large intrinsic dimension/non-rigid graphs

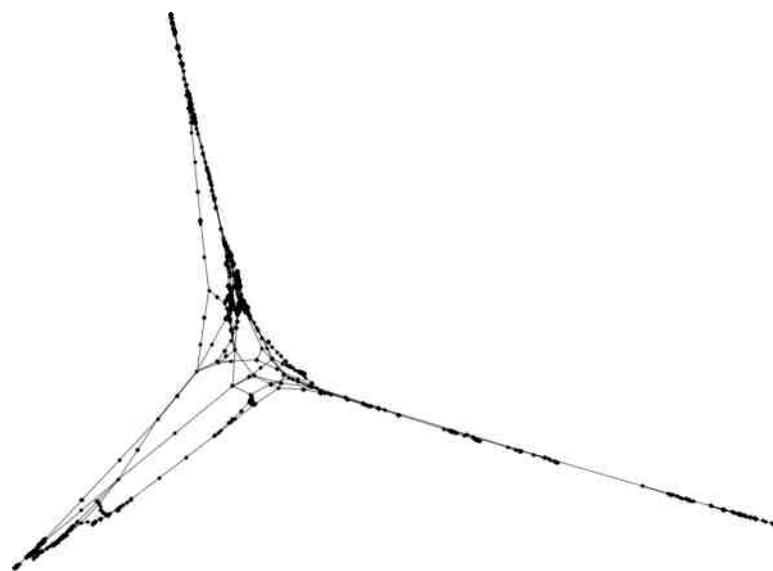
Drawing by some layout algorithms



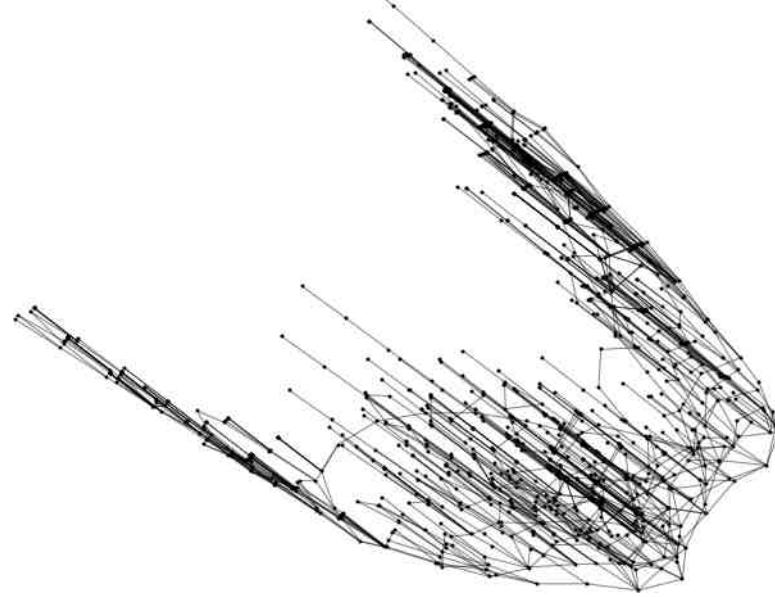
Spring (Stress) Model



Spring-electrical model



Eigenvector (Hall's) method



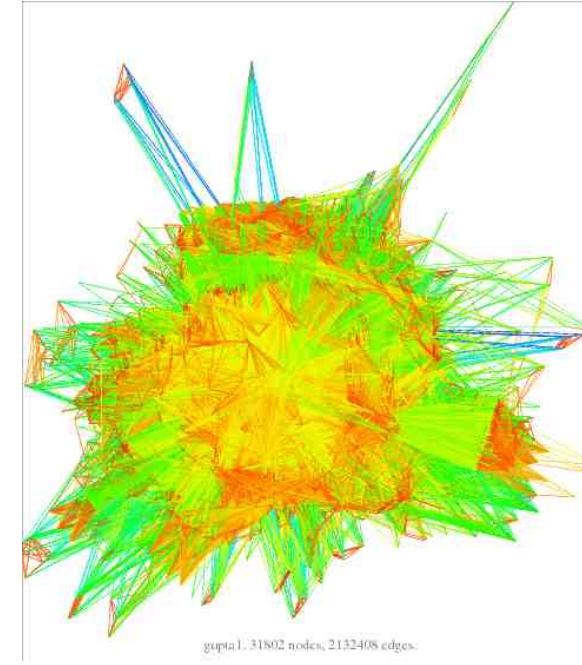
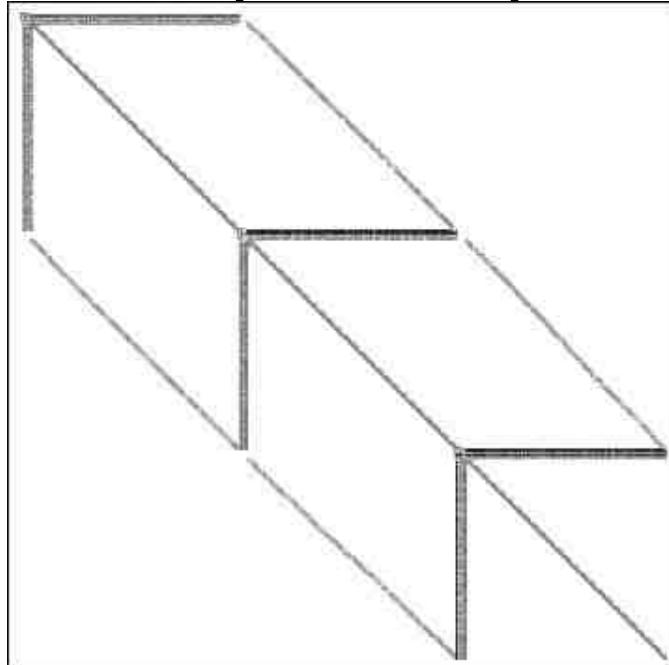
High dimensional embedding

Graph visualization: challenges

- Some graphs are difficult to layout
- Size of graphs get larger and larger
- Making complex relational data accessible to the general public
- Large graphs with predefined distance (can't use spring model)

Challenges: some graphs are hard

- Multilevel spring-electrical works for a large number of graphs, but not all!
- When applied to some real world graphs, the results: not good...
- Example: Gupta1 matrix. 31802×31802 .



Problem: Multilevel Coarsening

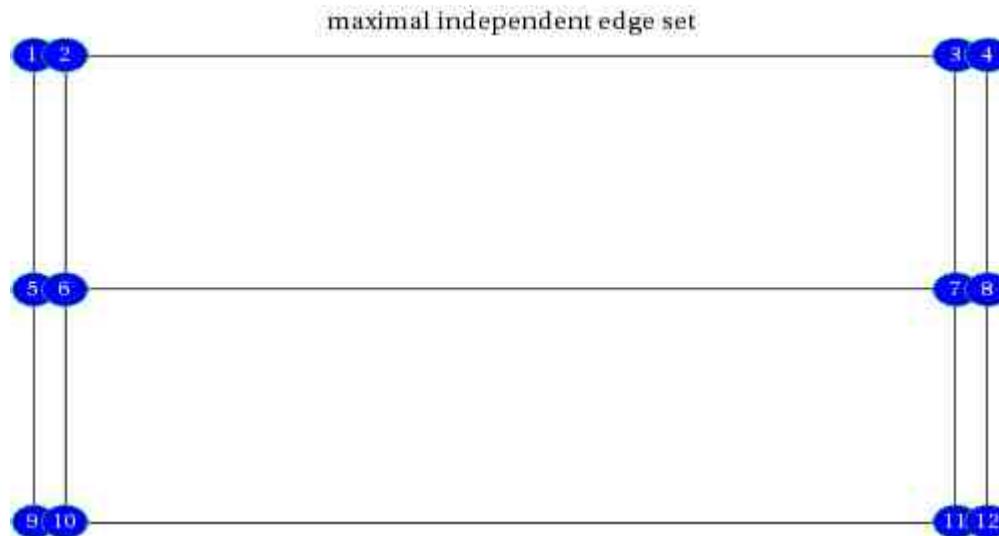
- A look at the multilevel process on Gupta1
- The problem: usual coarsening schemes do not work well

level	$ V $	$ E $
0	31802	2132408
1	20861	2076634
2	12034	1983352
3	11088	← Coarsening too slow, stop!

- Coarsening has to stop to avoid high complexity!

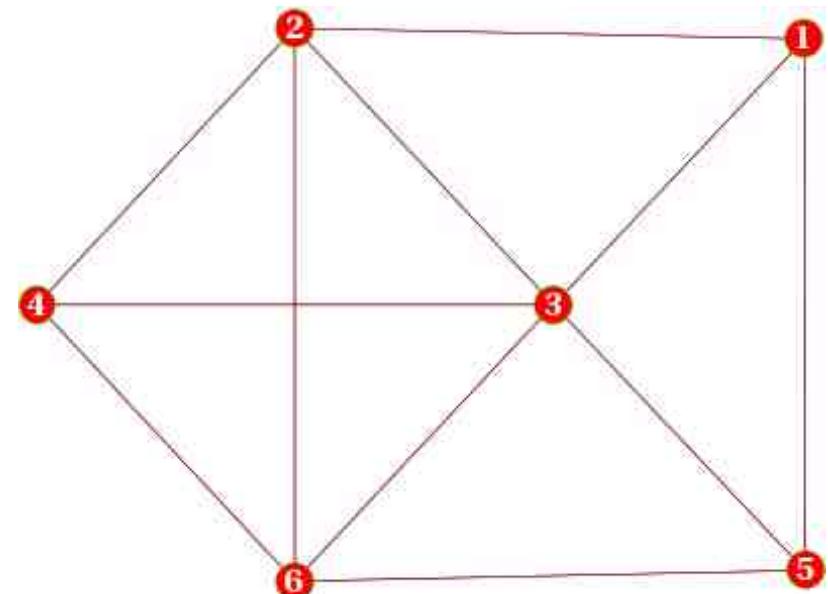
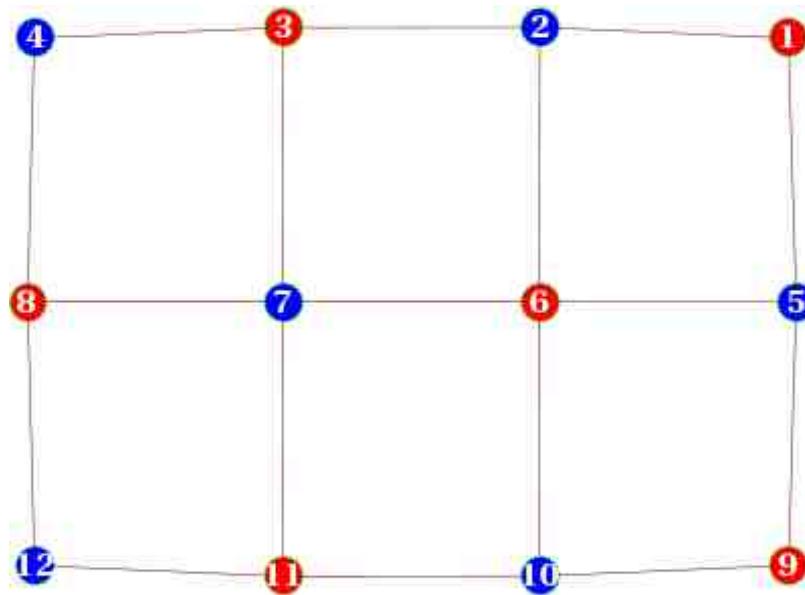
Multilevel Coarsening 1

- A popular coarsening scheme: contraction of a maximal independent edge set



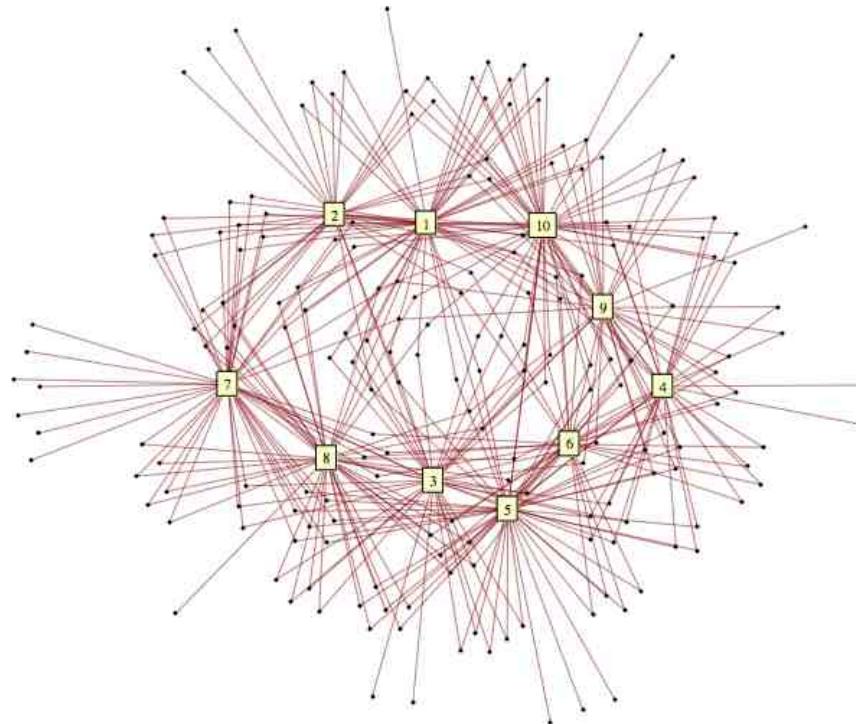
Multilevel Coarsening 2

- Another popular coarsening scheme: maximal Independent vertex set filtering



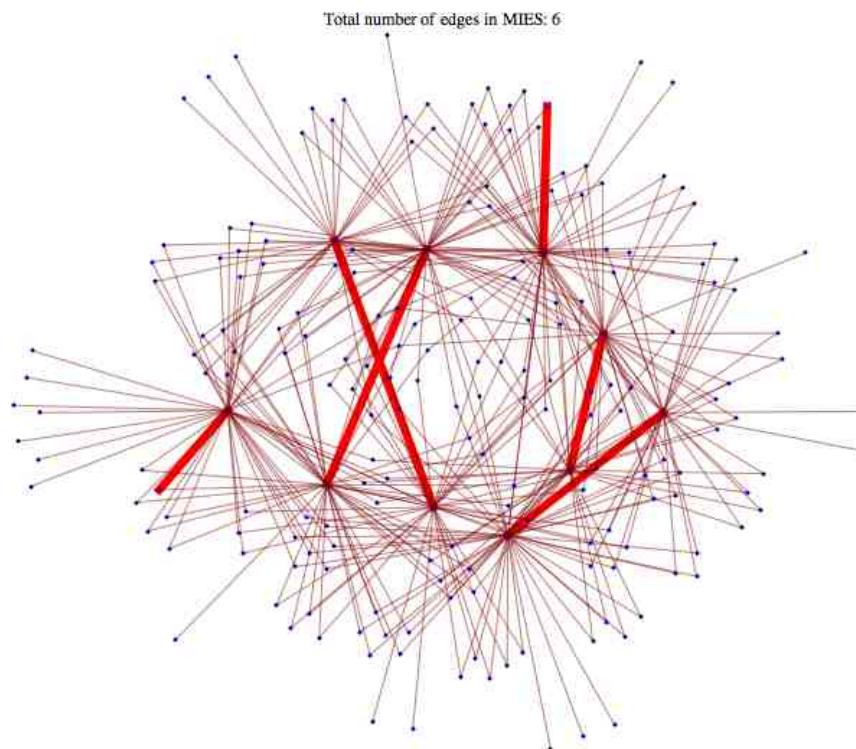
Coarsening Scheme Fails

- The usual coarsening algorithms fails on some graph structures
- Example: a graph with a few high degree nodes
- Such structure appears quite often in real world graphs



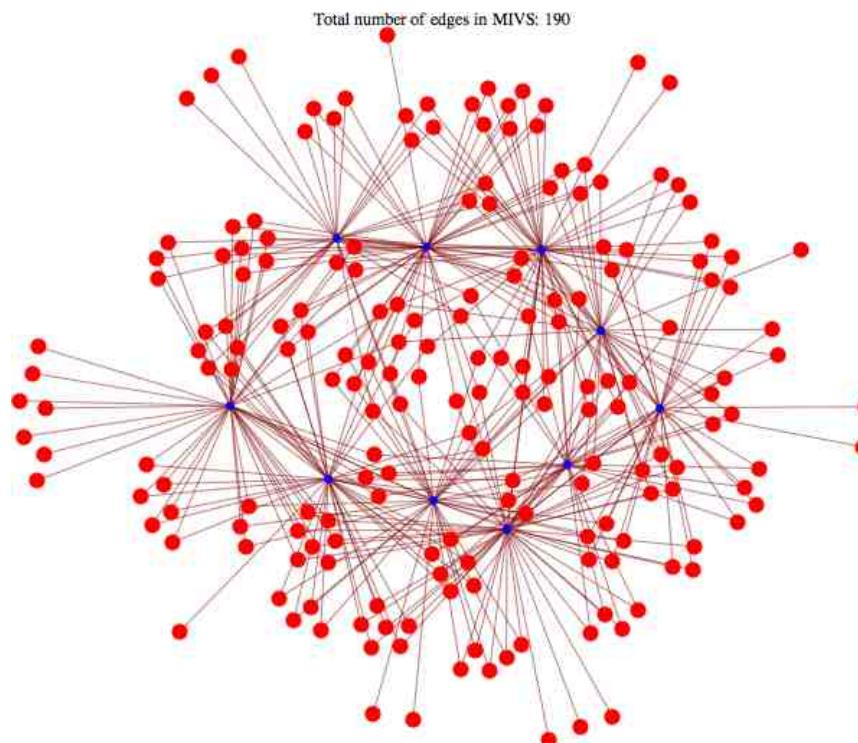
Coarsening Scheme Fails

- Maximal independent edge set coarsening: 6 edges out of 378 picked



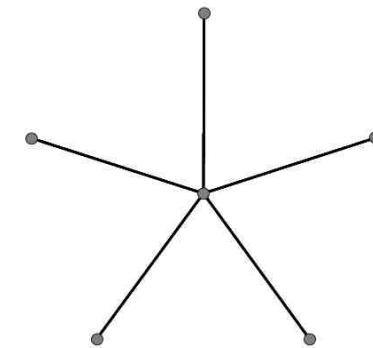
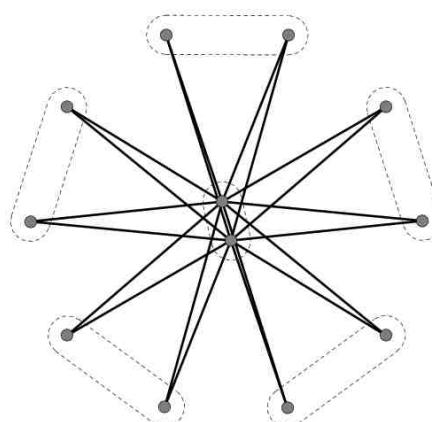
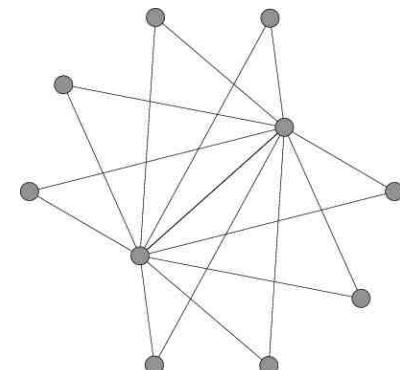
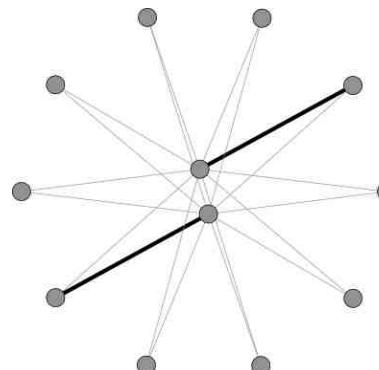
Coarsening Scheme Fails

- Maximal independent vertex set coarsening: all but 10 are chosen



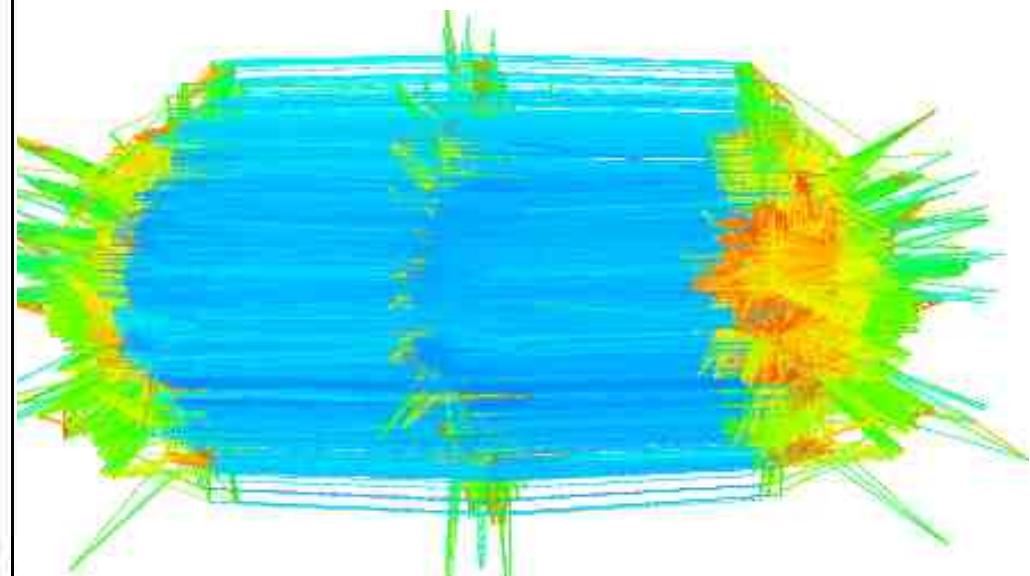
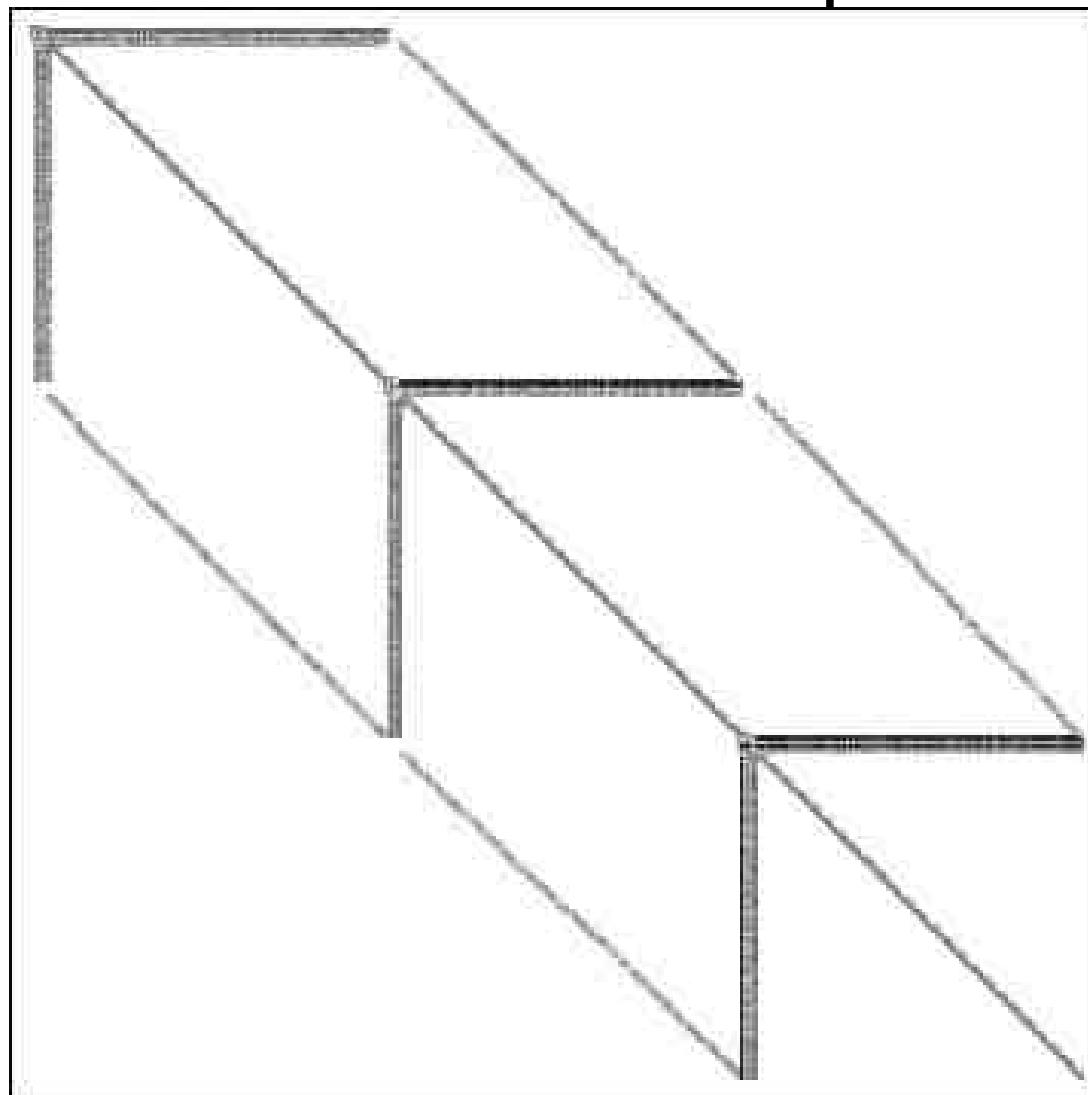
Better coarsening

- The solution: recognize such structure and group similar nodes first, before maximal independent edge/vertex set based coarsening.
- Instead of
- We do

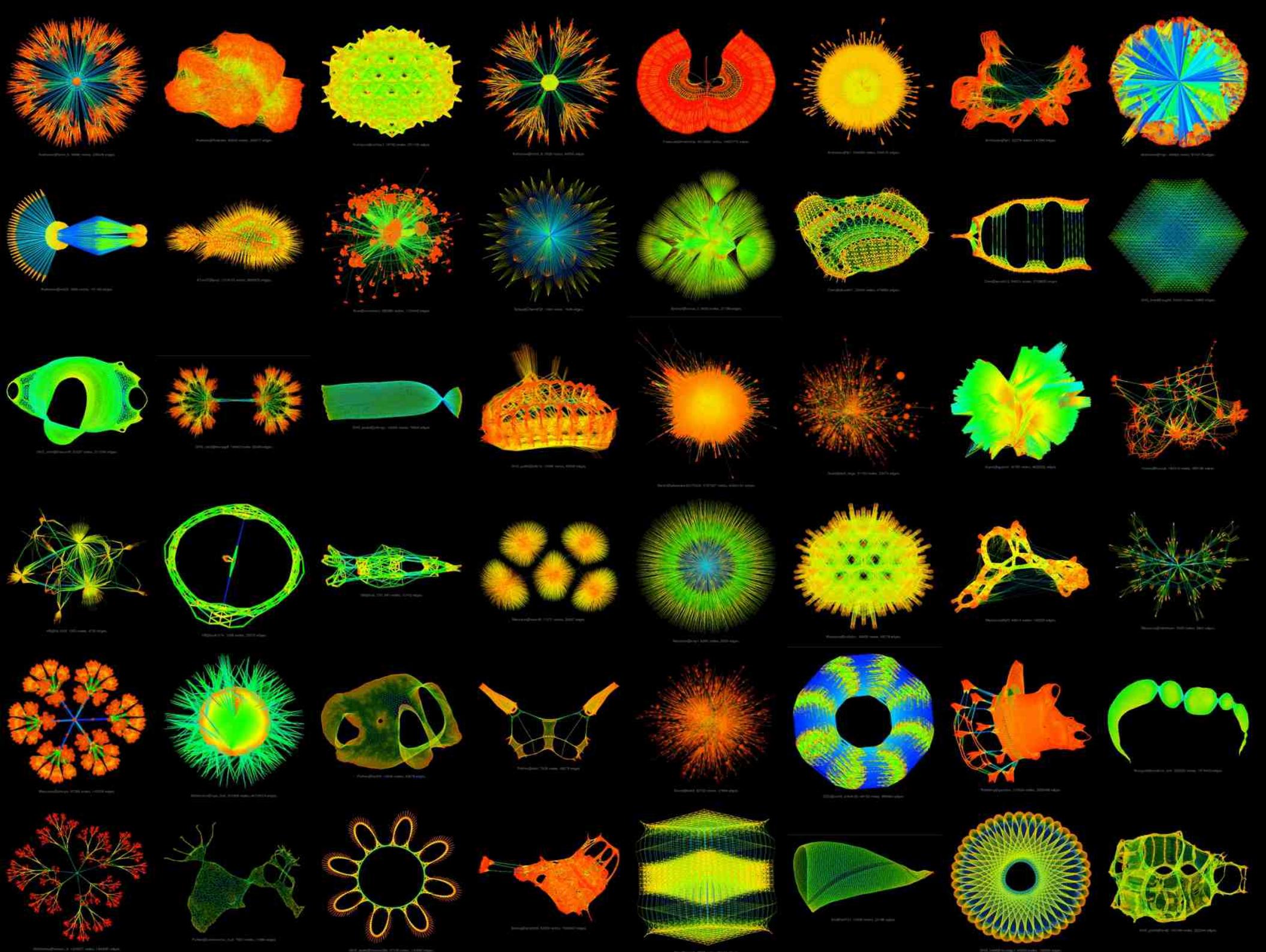


Better coarsening

- The result on Gupta1 matrix



gupta1, 31802 nodes, 213240K edges.

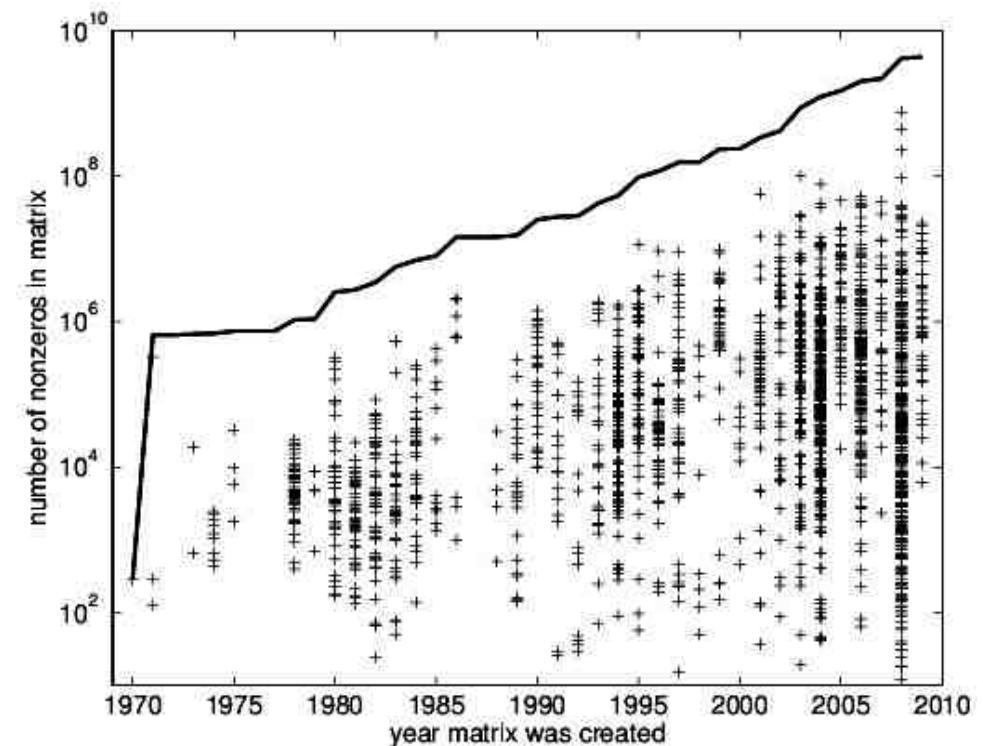
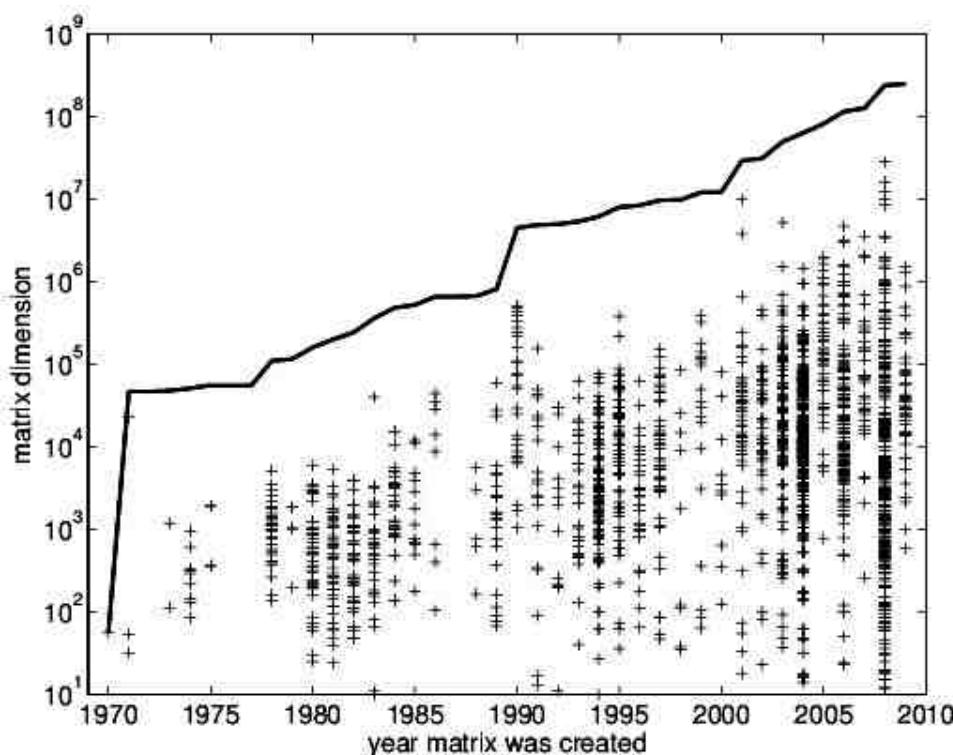


Challenges: size keeps increasing

- Many different types of matrices: a good testing ground for linear algebra/combinatorial algorithms
- E.g., testing on this collection revealed the coarsening issued discussed

Challenges: size keeps increasing

- Size keeps growing!
- Largest matrix: 50 million rows/columns and 2 billion nonzeros

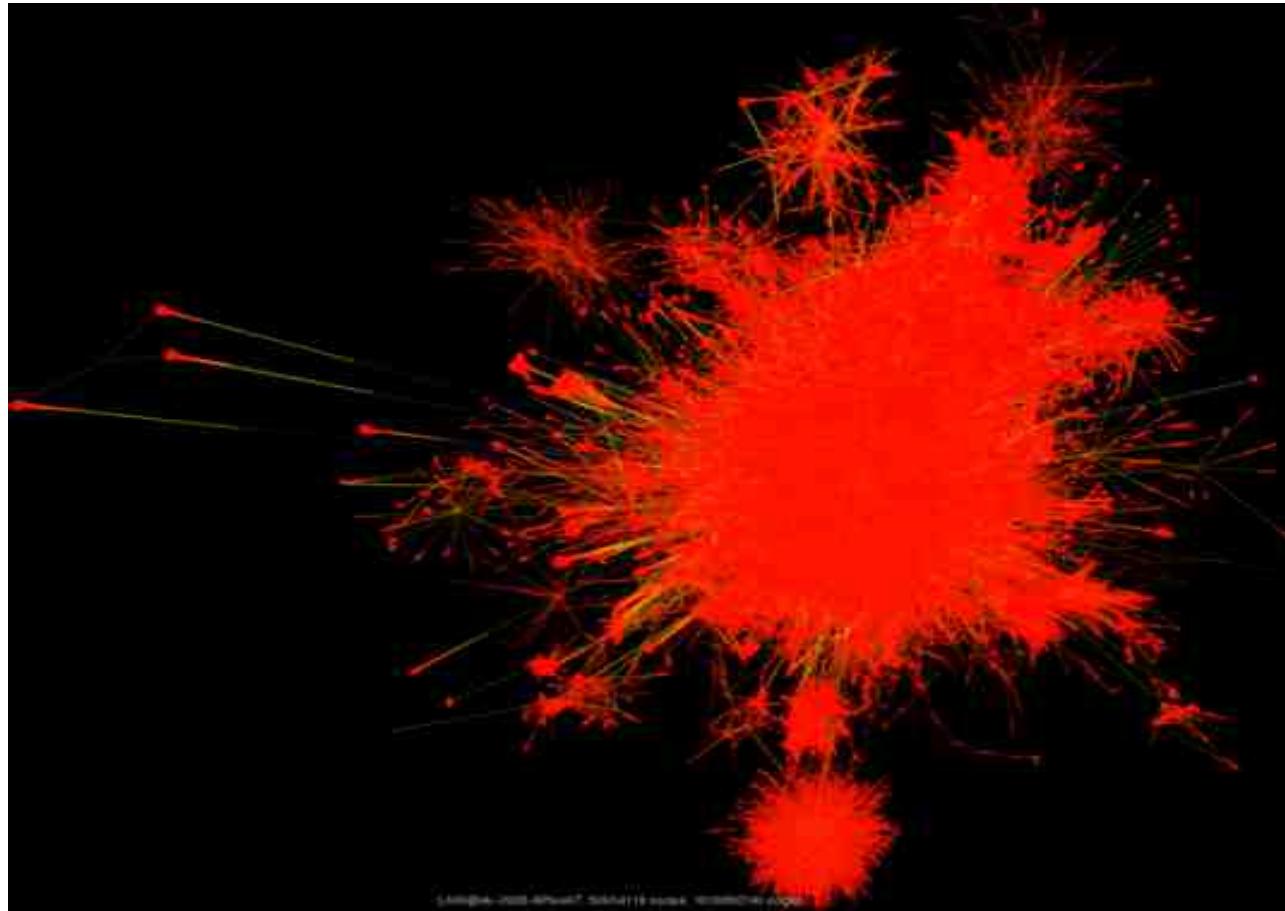


Challenges: size keeps increasing

- The largest graph: sk-2005, crawl of the .sk (Slovakian) domain
- 2 billion edges
- Challenge to layout: need 64 bit version.
- Challenge to rendering: 100 GB postscript.
- Convert to jpg/gif using ImageMagic: crash.
- Solution: rendering using OpenGL.
- But my desktop only has 12 GB → rendering in a streaming fashion (does not stores the edges).

The largest graph in the collection

- The result:



- Challenges: some graphs are hard to visualize
 - small world graph like that!

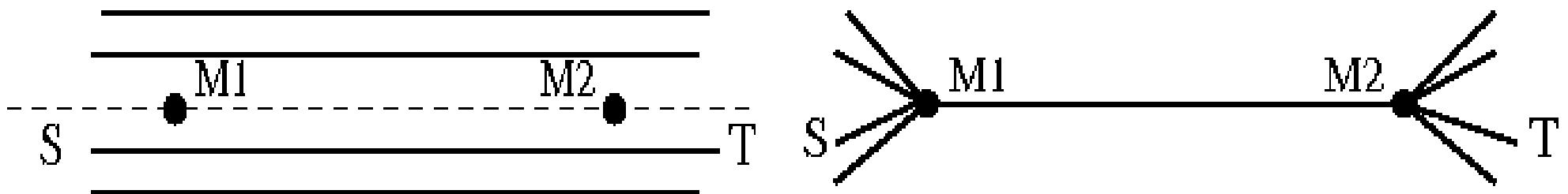
Challenges: hard graphs

- Visualizing small world graphs
- Possible tool: filtering. E.g., via k-core decom.



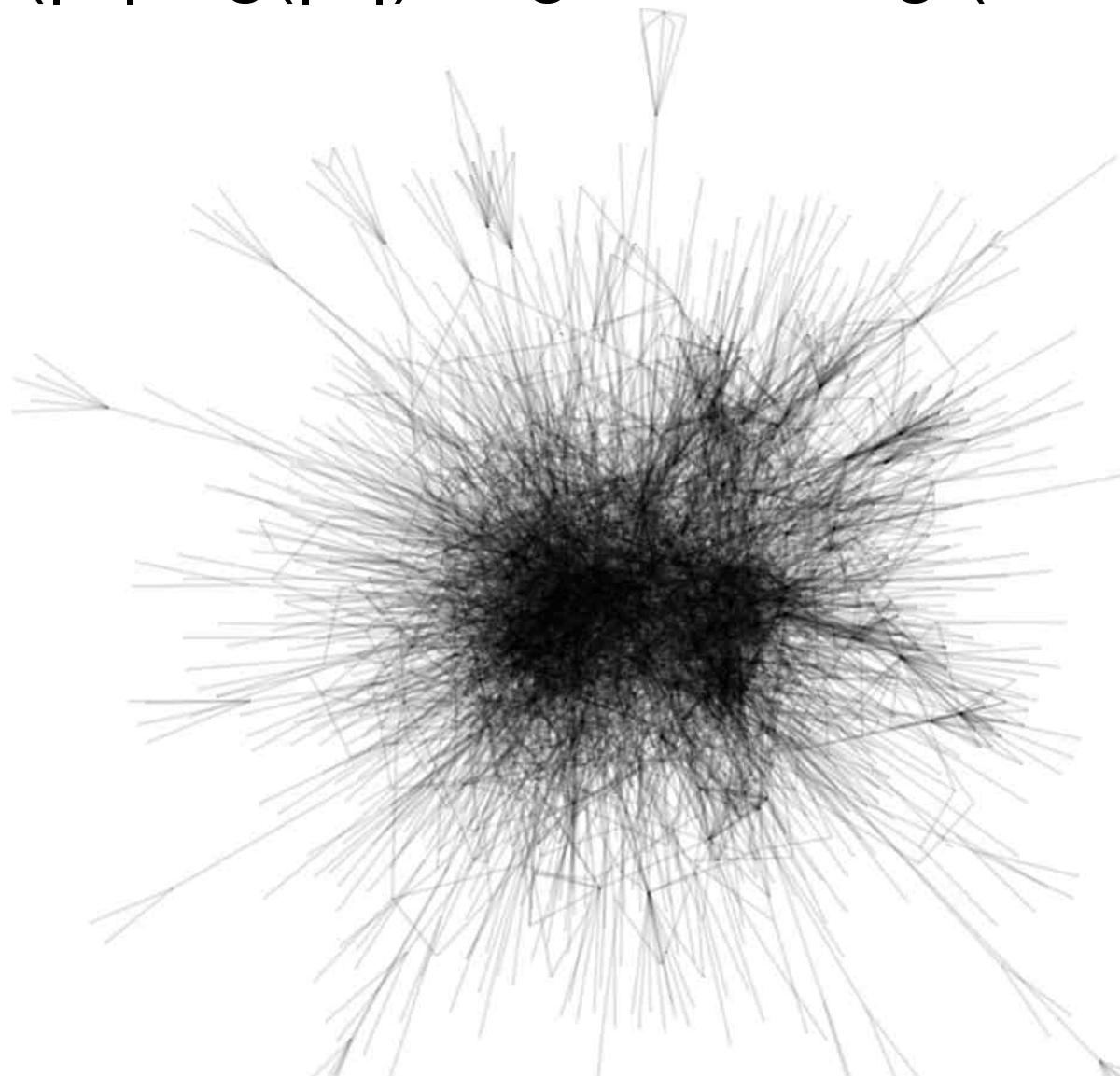
Challenges: hard graphs

- Visualizing small world graphs
- Possible tool:
 - abstraction (icons for cliques)
 - hierarchical (multilevel) view
 - fish-eye view
- Another possible tool: edge bundling



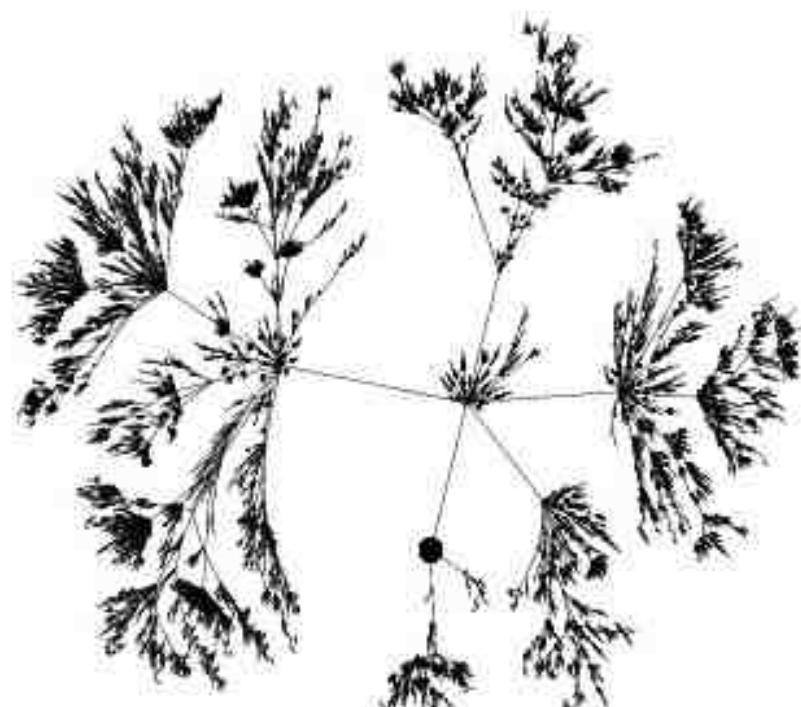
Challenges: hard graphs

- Fast $O(|E| \log(|E|))$ edge bundling (with Gansner)



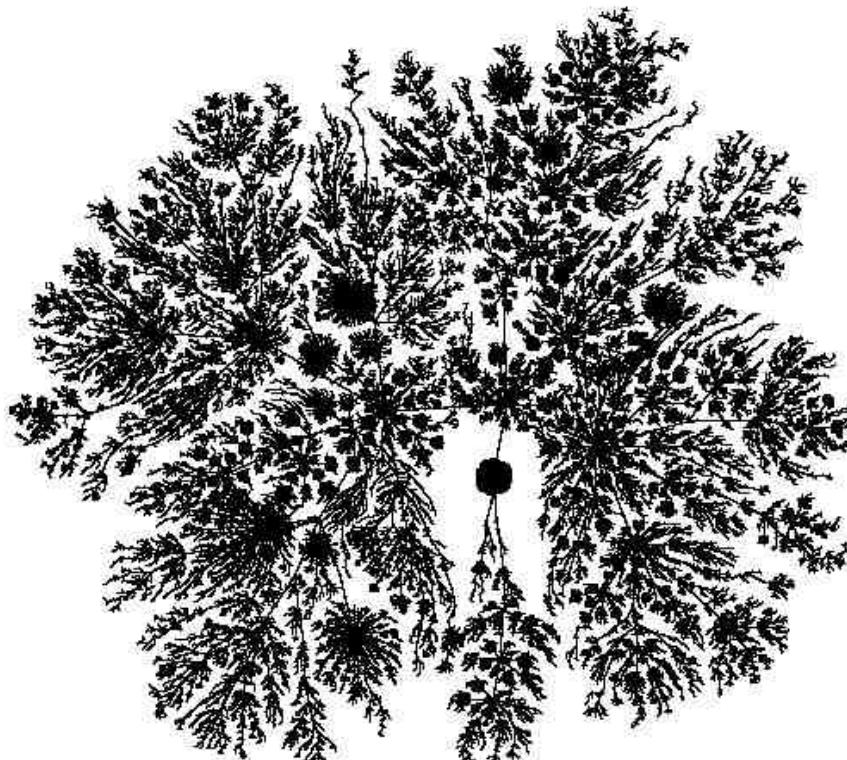
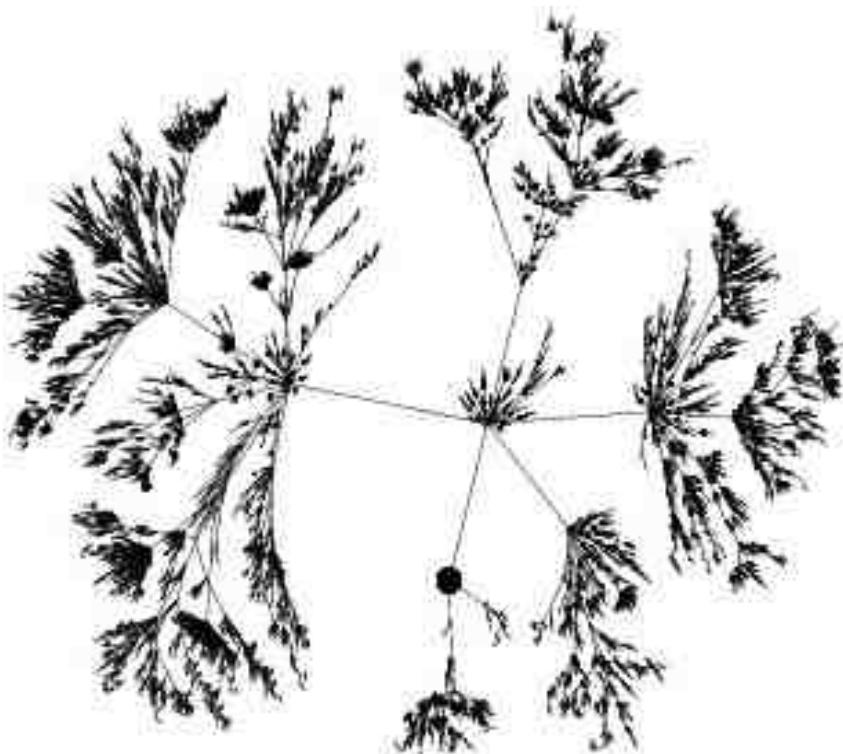
Challenges: some graphs are hard

- Even drawing trees can be tricky!
- Spring-electrical model suffers from a “warping effect”.
- A spanning tree from a web graph



Drawing trees

- Proximity stress model (with Koren, 2009)



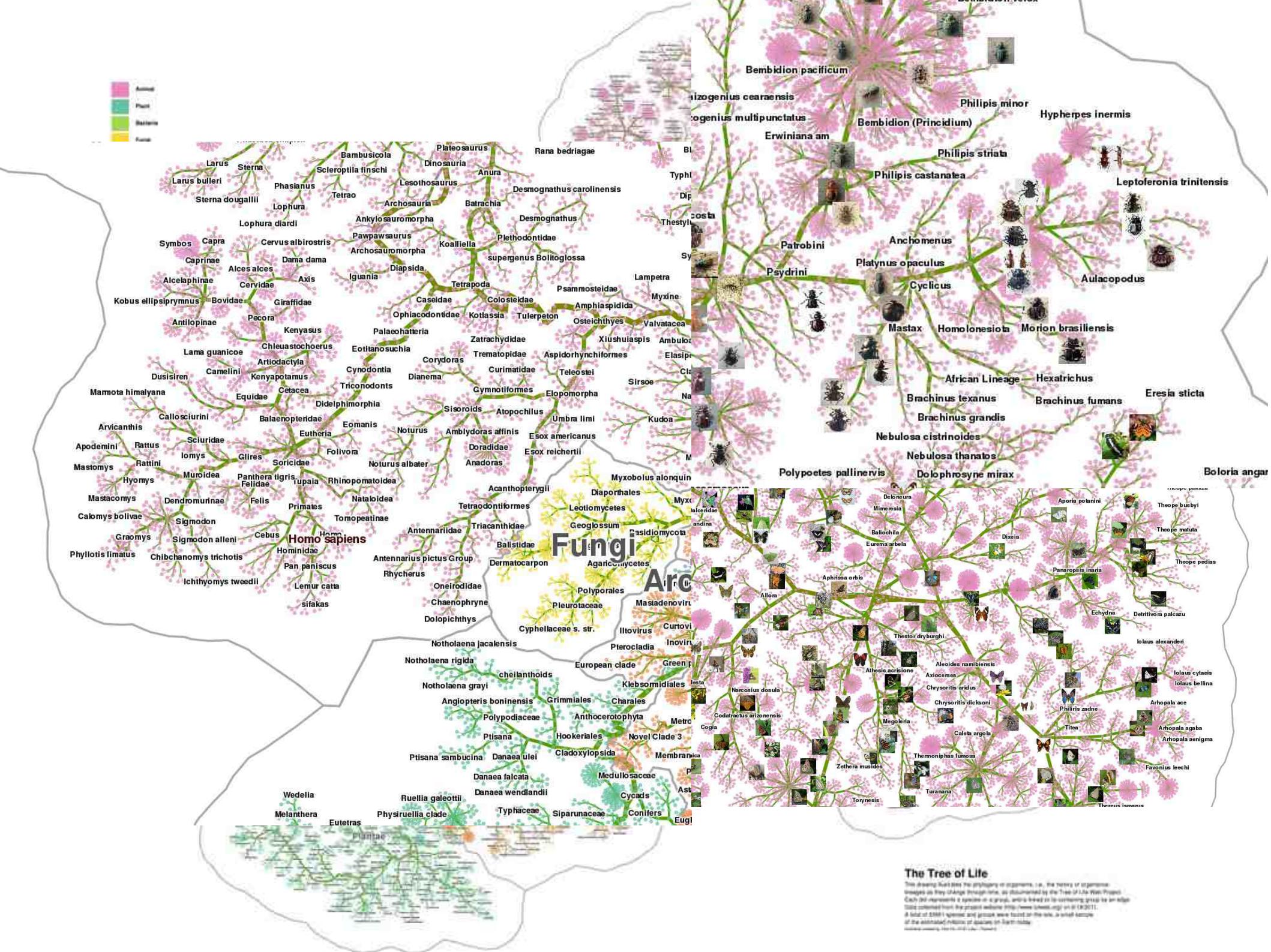
Animal
Plant
Bacteria
Fungi

Fungi

ARC

The Tree of Life

This drawing illustrates the phylogeny of organisms, i.e., the history of life as we know it through time, as documented by the Tree of Life Web Project. Each clade represents a species or a group, and is linked to its containing group by an orange line. Data collected from the project website (<http://www.tolweb.org>) on 11 October 2011. A total of 331,113 species and groups were found on the site, which amounts to an estimated million of species on Earth.



An Internet map: Reagan/Dulles



Visualizing graphs as maps

- So far graphs → node-link diagrams
- Not familiar to the general public
- Example

Recommender System Visualization

- AT&T provides digital TV (U-verse).
- A few hundred channels: need a recom. system!
- Recommending TV shows
 - If you like X, you will also like Y & Z.
 - Based on SVD/kNN: similarity of shows
- Like to visualize to see if model makes sense
- Also provide a way for users to explore the TV landscape.

Recd

This figure is a network graph illustrating the relationships between numerous TV shows, movies, and other media entities. The nodes are represented by labels in bold black font, and the connections are shown as thin grey lines. The network is highly interconnected, with many nodes having multiple links to others.

The nodes include:

- Shows: Go, Dora, Wonder Pets!, The Backyardigans, Max and Ruby, Wow Wow Wubbzy, Yo Gabba Gabba!, Little Bill, Pinky Dinky Doo, Blue's Clues, Oswald, Miss Spider's Sunny Patch Friends, LazyTown, Toot & Puddle, Jack's Big Music Show, Franklin, Disney's Mickey Mouse Clubhouse, Handy Manny, Little Einsteins, Charlie &, Da Vinci's Inquest, Snapped, Match Game, Star Trek: The Next Generation, Deadliest Catch, Ghost Whisperer, Rules of Engagement, The Big Bang Theory, Smallville, net Earth, MythBusters, Survivorman, Dirty Jobs, Modern Marvels, The Dead Zone, Destroyed in Seconds, Treasure Quest, How Do They Do It?, Cash Cab, Overhaulin', Explorer, Naked Science, an Chopper, and Workin' Nation with.
- Movies: 64, Zoo Lane, Maurice Sendak's Little Bear, Maggie and the Ferocious Beast, and Blue's Clues.
- Other entities: The Upside Down Show, Little Einsteins, Charlie &, Da Vinci's Inquest, Snapped, Star Trek: The Next Generation, net Earth, The Dead Zone, an Chopper, Workin' Nation with, and several network logos and names.

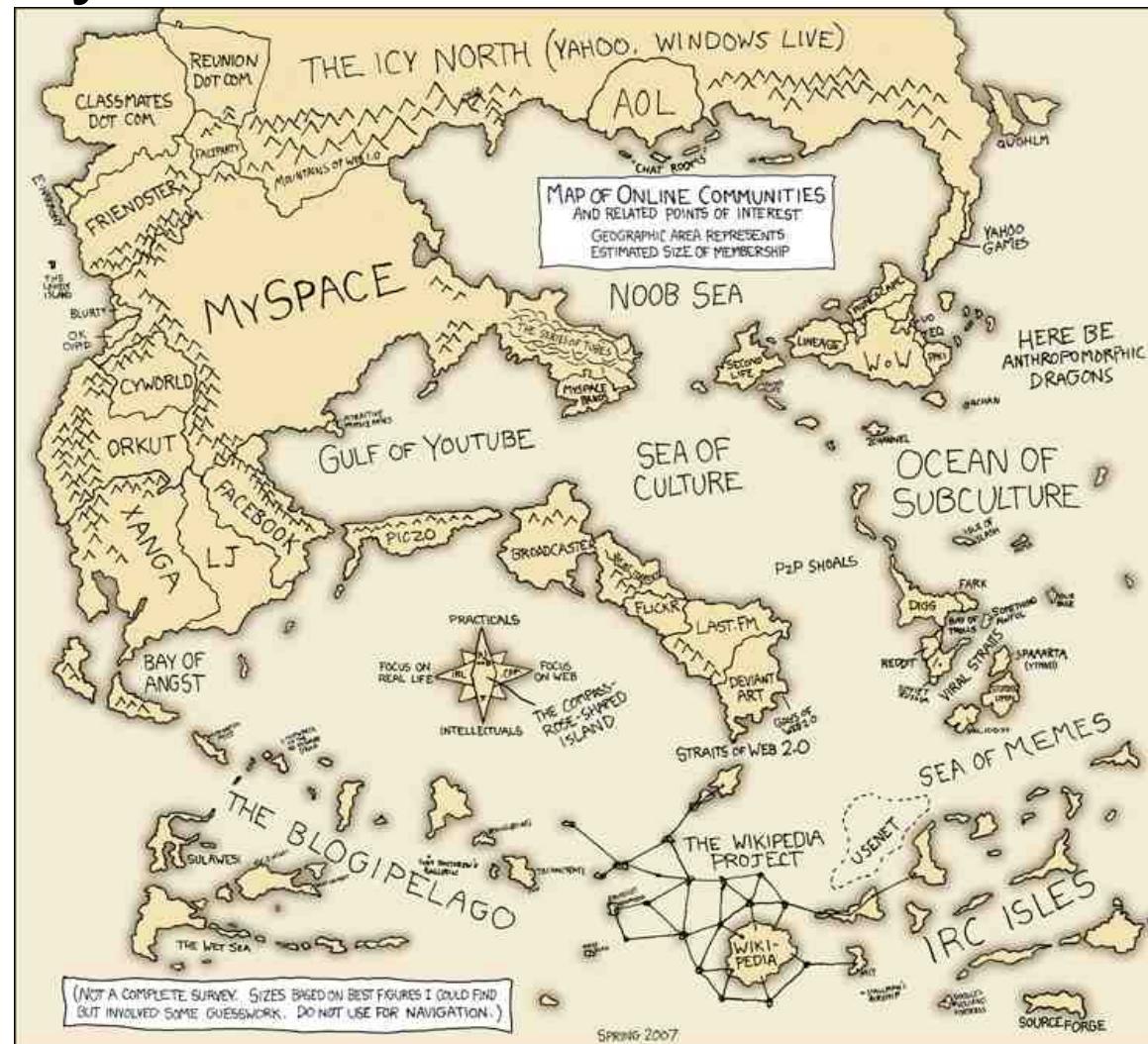
Recommender System Visualization

- Messy. Not easy to understand for general public.

Better defined boundary → a map?

Recommender System Visualization

- Virtual maps are used frequently
- E.g., “online community”, circa 2007
- Can we make a map like that, but use real data?

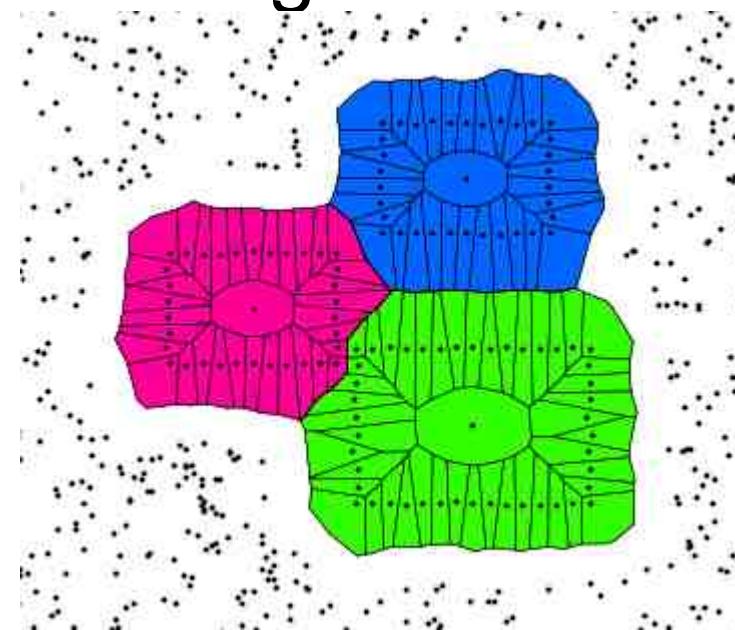
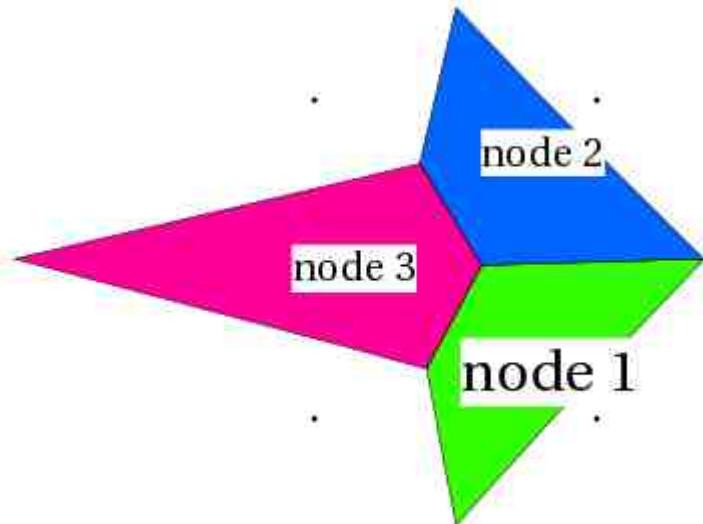


Gmap algorithm

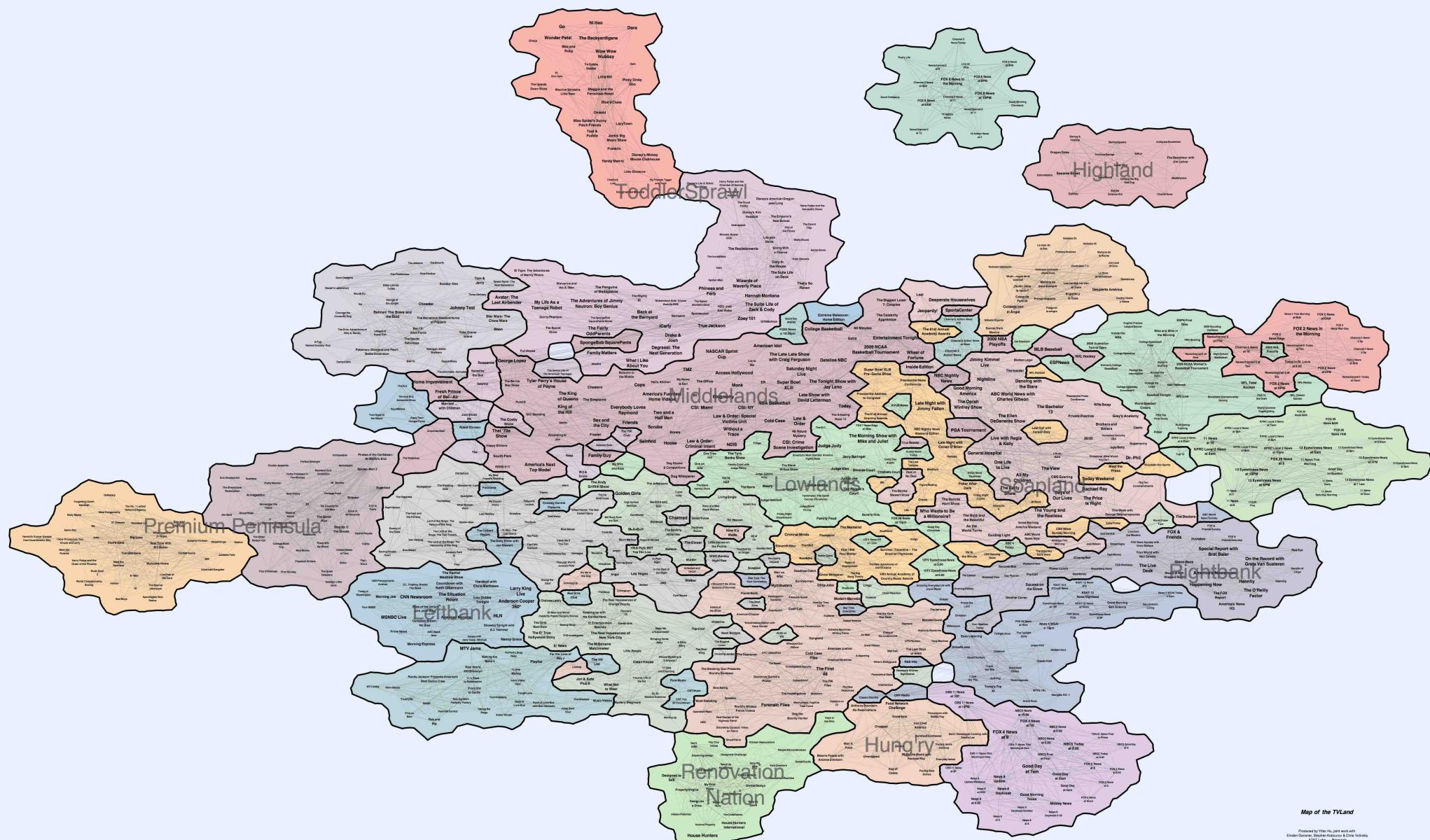
- Gmap algorithm (Gansner, Hu & Kobourov, 2010) – available as *gvmap* from GraphViz.
- Four step process
 - embedding
 - clustering
 - mapping
 - coloring

Gmap algorithm

- Embedding + clustering use standard algorithm
- Mapping. Based on Voronoi diagram



Gmap algorithm

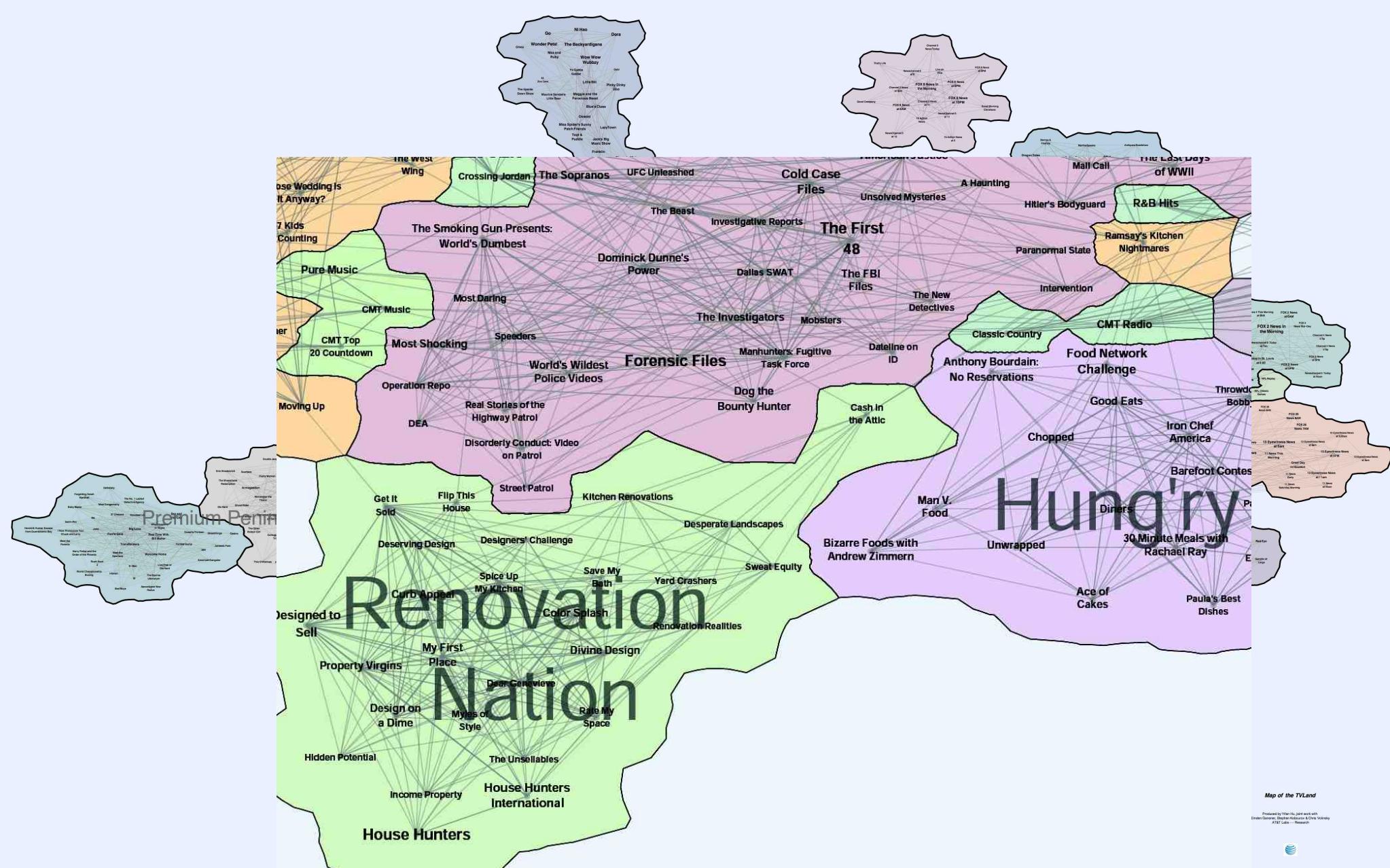


But the coloring needs improvement!

Gmap algorithm

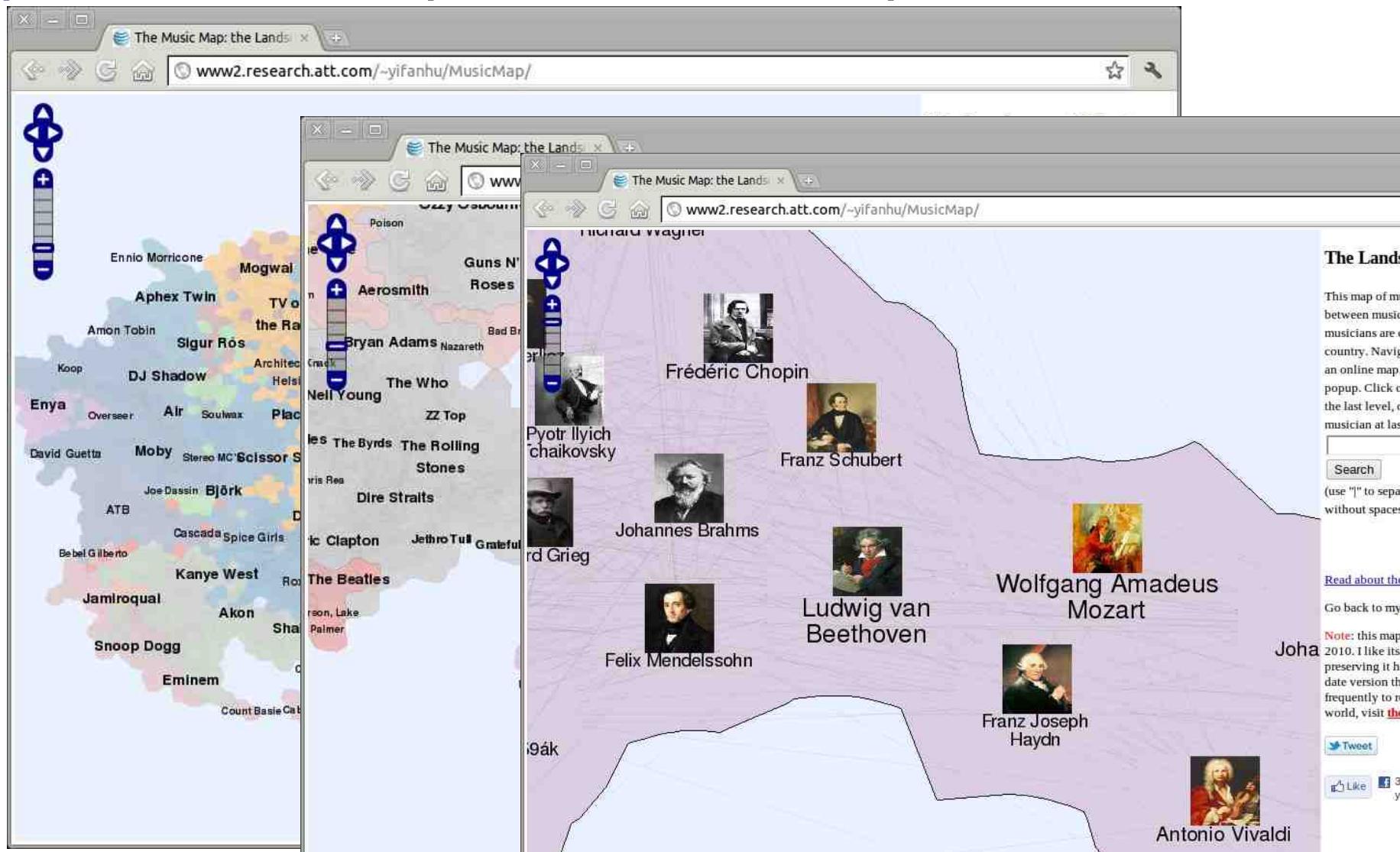
- Coloring algorithm: maximize difference between neighboring countries.
- Solution: solve a graph optimization problem.
- Also known as the anti-bandwidth problem.
- Final result:

Gmap algorithm



Gmap applied to other areas

- Map of music; map of movies; map of books etc



Twitter Visualization



What are people talking about wrt the topic “news”?

#pharma news: ACT Announces Second Patient with Dry AMD Treated in U.S. Clinical Trial with RPE Cells Derived from ...
<http://t.co/EsqBjL00>

Nashville News Home Destroyed, Two Others Damaged By Fire: NASHVILLE, Tenn. A home was destroyed and two neighbo...
<http://t.co/dcxUF7nO>

Danielle woke me up to the GREATEST news ðŸ˜·.

RT @lbaraldo: devo dire che l'app #fineco e' quasi meglio del sito. I grafici immediati di alcune aree sono spettacolari e le news sono ...

The Affiliate Networks - DE News wurde gerade verÃ¶ffentlicht! <http://t.co/RbOt8OtJ> –, Topthemen heute von @tddepromotions @affilinet_news

@jsimoniti I saw it on the news and could tell fairly easily

RT @The1Daily: That feeling when your friends try to tell you 1D news & you're like "I already know. Get on my level, dude. PROUD Direct ...

Valerio Pellegrini Digital News is out! <http://t.co/UZacEO9k> –, Top stories today via @palettod @dr8bit @alldigitalexpo @ggrch In the news: (Examiner) Fake AT&T bills being used to deliver malware: <http://t.co/IWWtfhec>

[NEWS PIC] 120416 Kangin's comeback - Happy Kyuhyun :D <http://t.co/X1J1djam>

RT @SizzlinStockPix: STOCKGOODIES PLAYS OF THE WEEK: \$STKO news just out link below <http://t.co/FEYe2TR0>

@NatashaSade_ GM homegirl..... We have until tomm to file..... I just seen it on the news lol FYI

My horoscope said don't worry about it.. I just news to find something to do with my time to get my mind off of it

RT @Real_Chichihu: SM should release news to slap that stupid official from that stupid music site

Ball State Daily News: Speaker informs students about female genital mutilation - <http://t.co/FuN5LqKo> via <http://t.co/rkaZhaCv>

Twitter Visualization

- Browsing can be tedious
- May even misses the overall picture
- Characteristics of Twitter stream
 - very short text (140 char)
 - streaming (3,000 tweets per second. 6X 2010)
 - considerable cross-copying (RT) and spontaneity
- What we like to see:
 - A “big picture” view
 - Clustered and summarized
 - Detail on demand

Twitter Visualization

- The approach we propose: a succinct high level visual clustering, with textual summary, and details on demand
- We will visualize only tweets relating to a keyword of interest

Tweet Similarity

- Finding similarity of tweets
 - either LDA, which gives distribution of topics over words, then document over topic. Then similarity based on topic distribution
 - or, treat each tweet as a vector of words, scaled using tf-idf. Followed by cosine similarity

$$\text{tf-idf}(t, d) = |\{t|t \in d\}| \times \ln \frac{|D|}{|\{d|t \in d \text{ and } d \in D\}|}$$

- We found that for tweets, the simpler tf-idf based similarity works just as well

Tweet Similarity

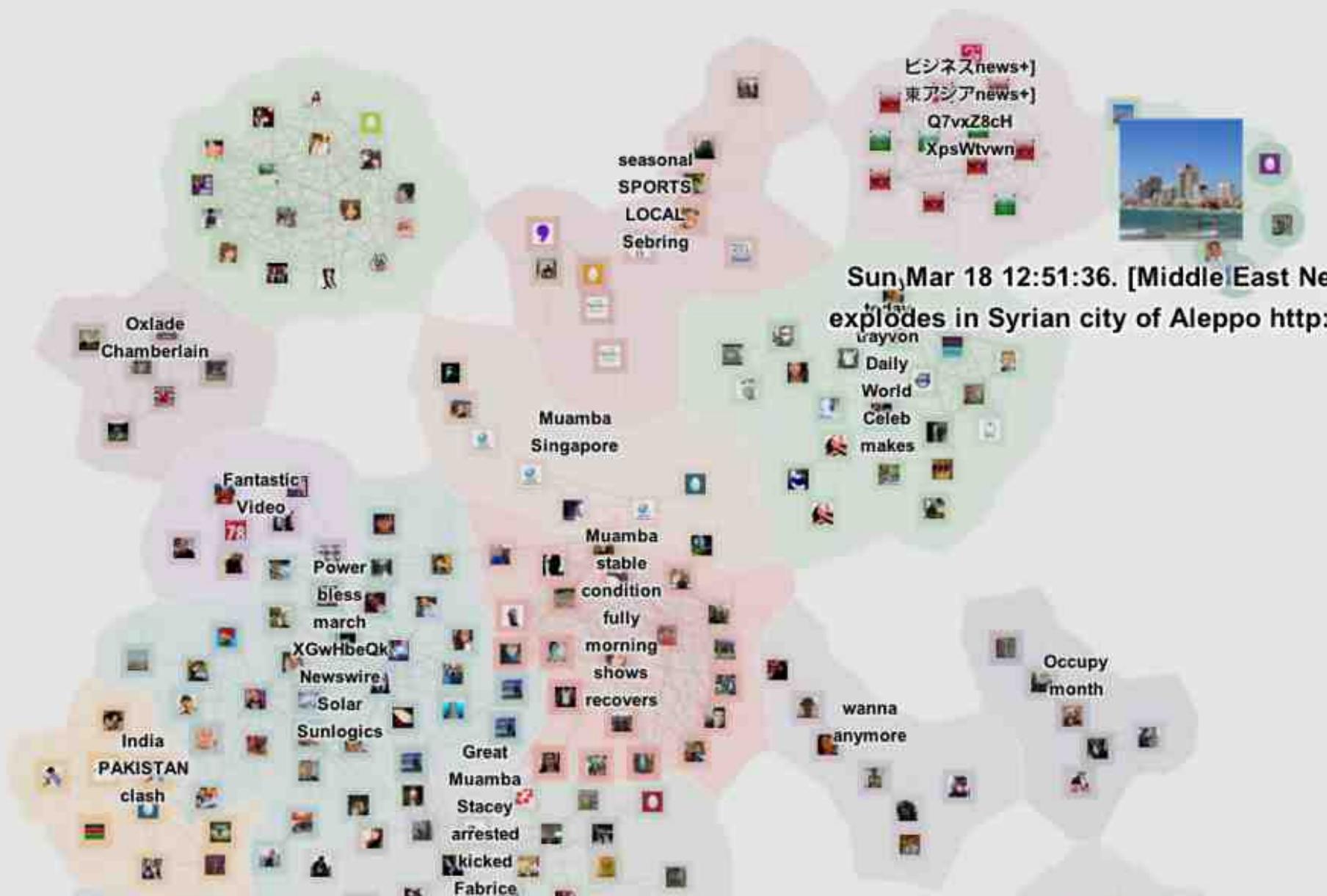
tibesti.research.att.com/twitterscape/news/

[Google](#)  [More](#)  [List Candidates](#)  [pkuvis: Weibo View](#)

SUN MAR 10 14:31:42 2013 [REDACTED] PREVIOUS PAGE PREVIOUS WEEKS

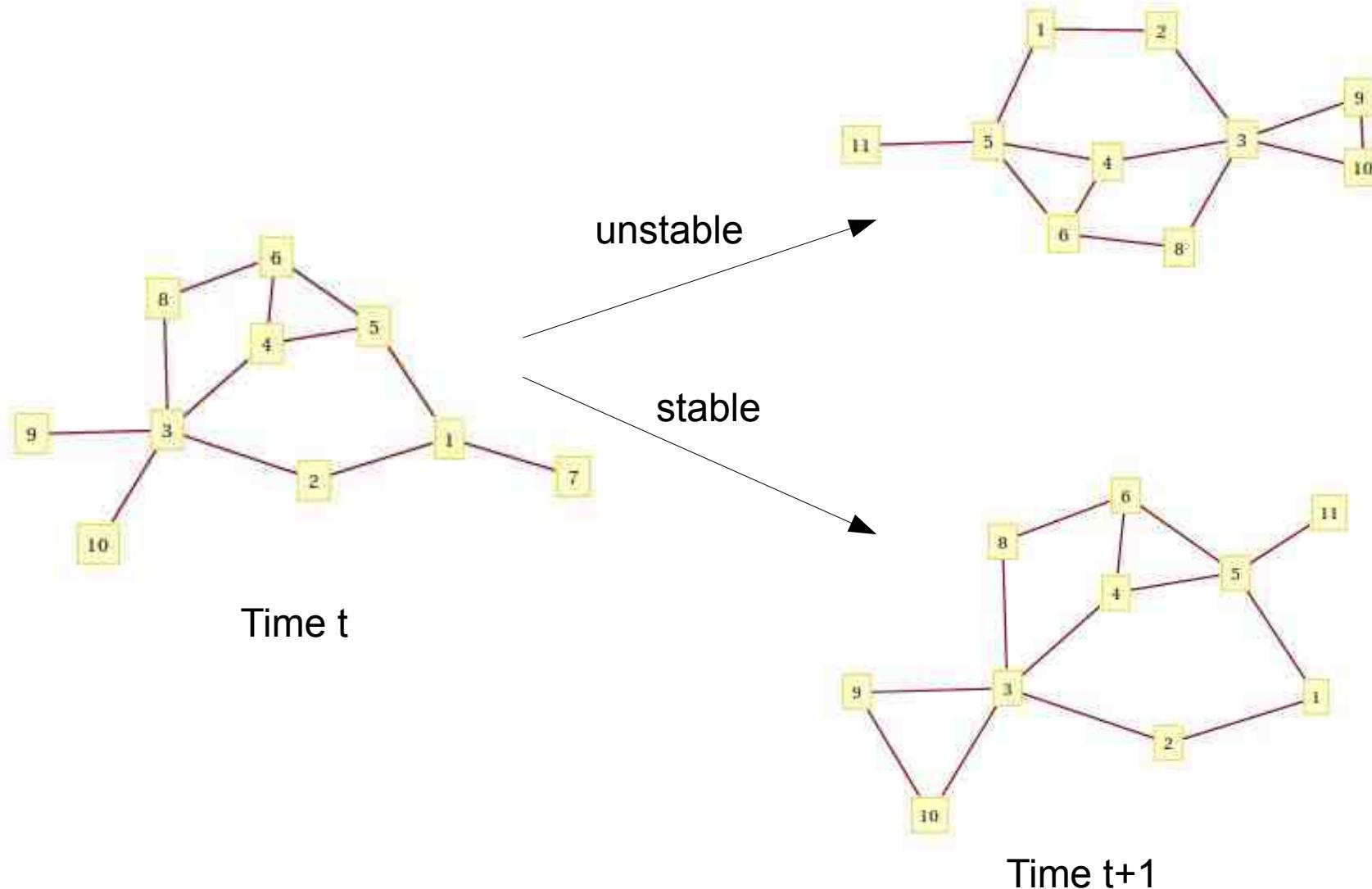
Sun Mar 18 12:51:25 Mauu RT @fellifellie RT @YeppopoKPOP: (News/Video) Seungri Mengajari Fans Menari • Wiper Dance’ Dari Lagu ... http://t.co/kWtsZ7Rg

Sun Mar 18 12:51:25 Lots of great guests on @anhqdcc @HaleyBarbour @SenatorSessions Puerto Rico Gov. Romney Sun Mar 19 12:51:24 Dearie McDearie - GMW Neighborhoods - The Bonus Show w/ J. Frank Morris Indiana and ZH VOS - with Rommel



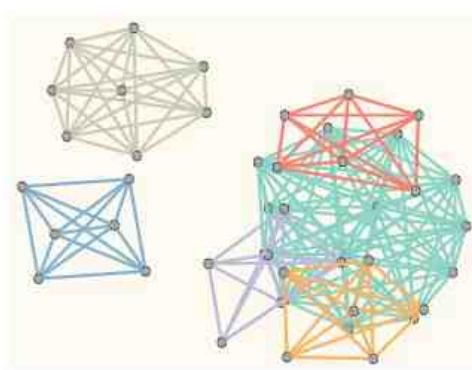
Dynamic Stability

- We ensure *layout stability* by warm start + Procrustes transformation



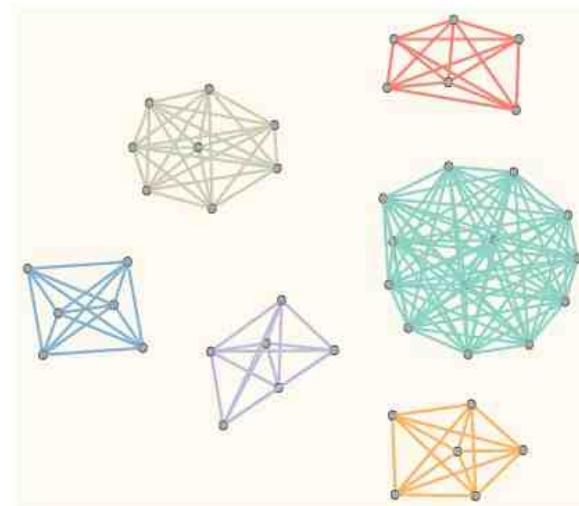
Dynamic Stability

- Component packing stability
 - disconnected component needs repacking stably



→

Rpack stably

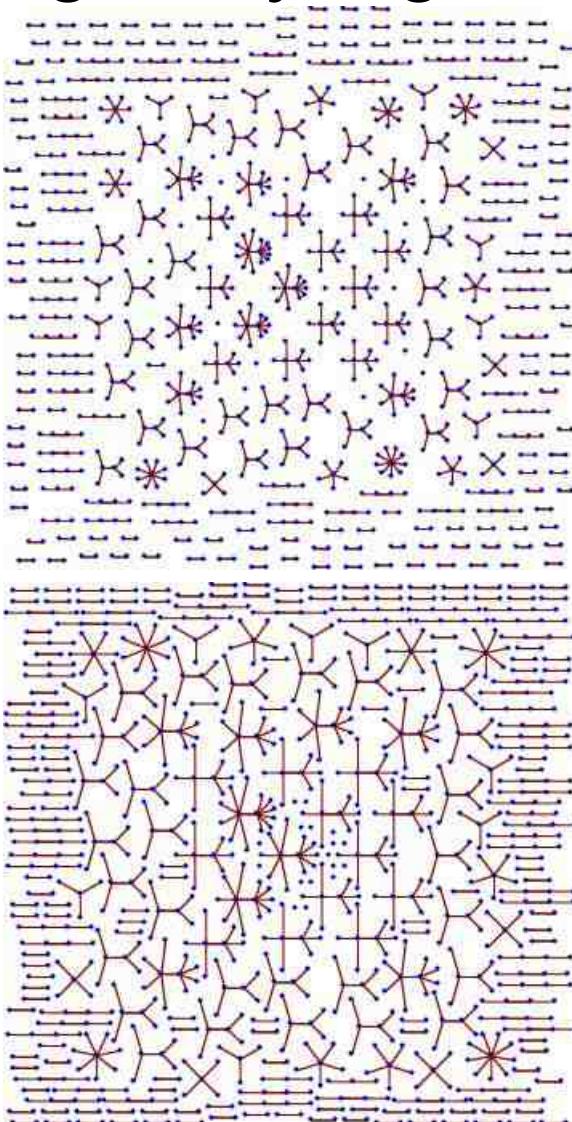


Dynamic Stability

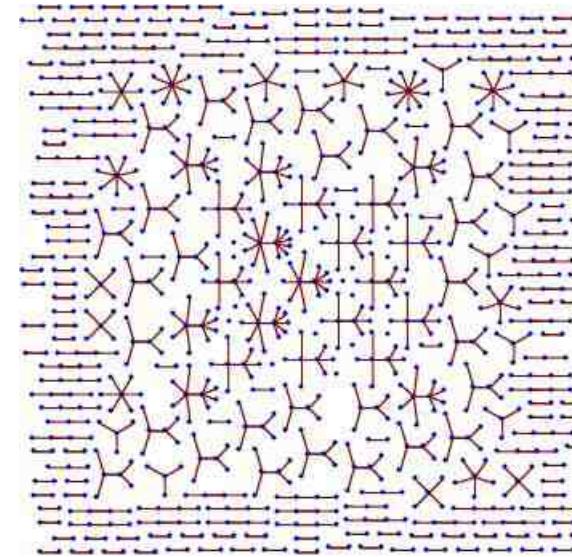
- Traditional packing algorithm: polyomino based greedy algorithm
 - Place the largest component at the origin
 - Place the next component as close to the origin as possible without overlap
 - repeat
- Can pack very tight

Polyomino-based Packing

- Traditional packing algorithm: polyomino based greedy algorithm. Good/tight packing

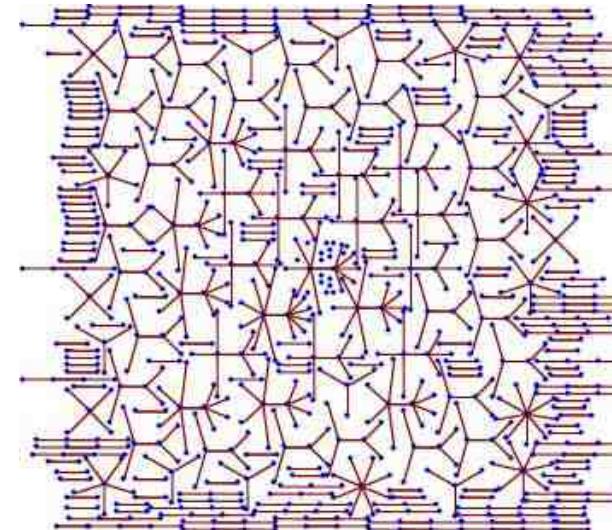


padding=10



padding=5

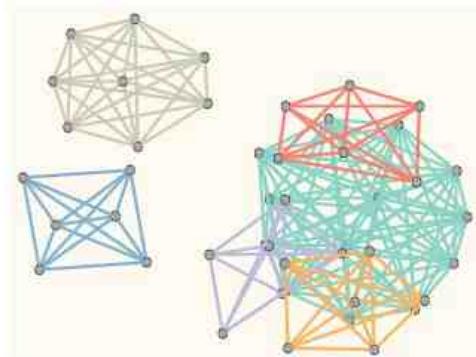
padding=3



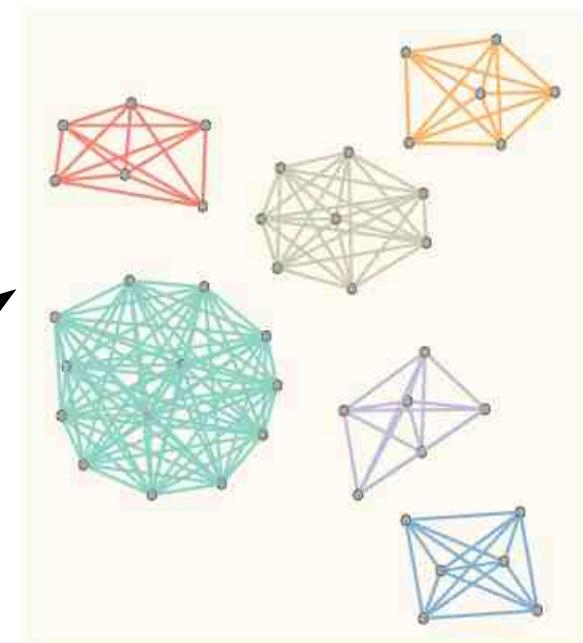
padding=1

Stable Packing

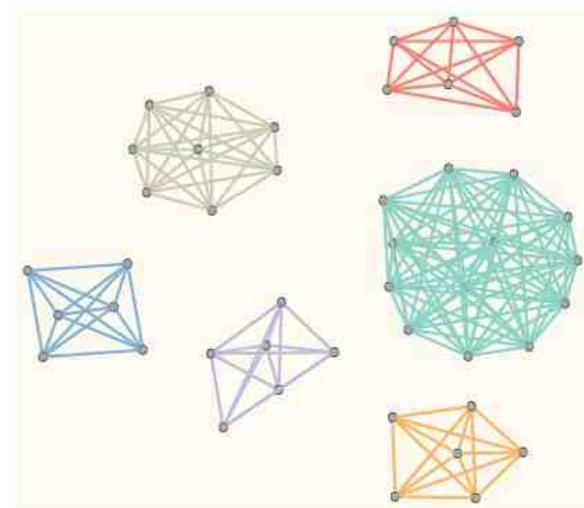
- Tradition packing pays no consideration to stability



Normal
Packing alg.

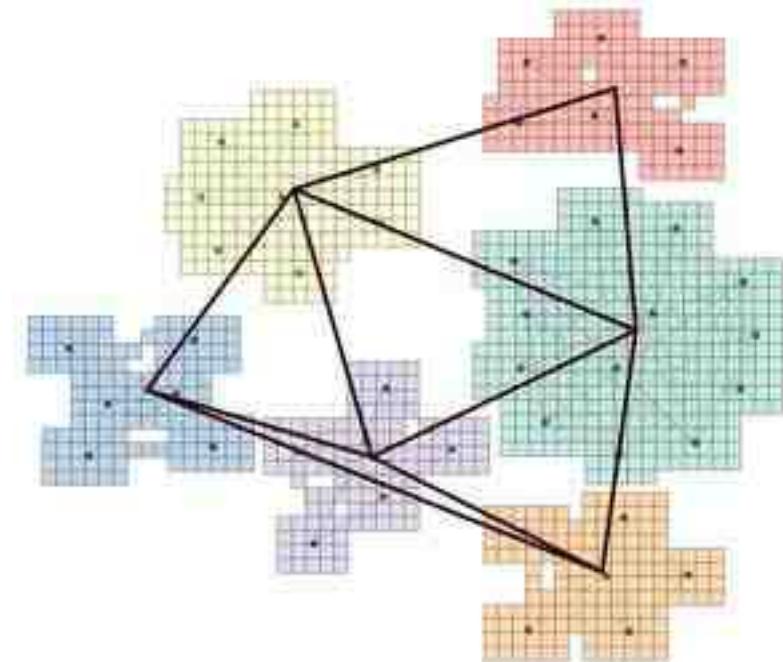
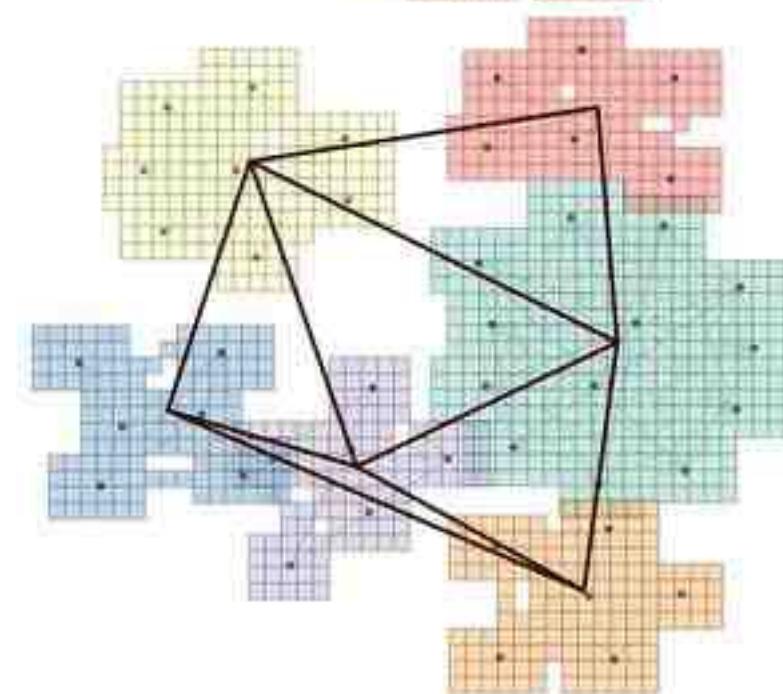
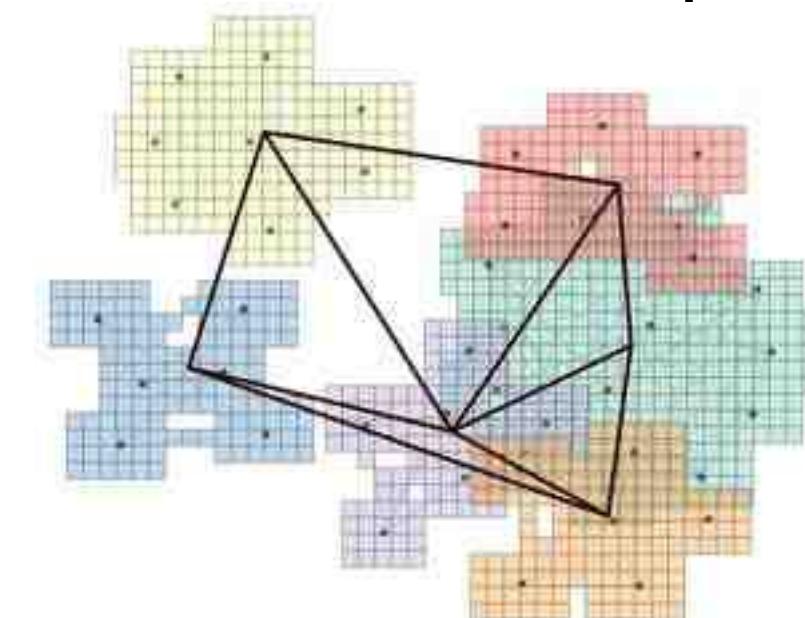
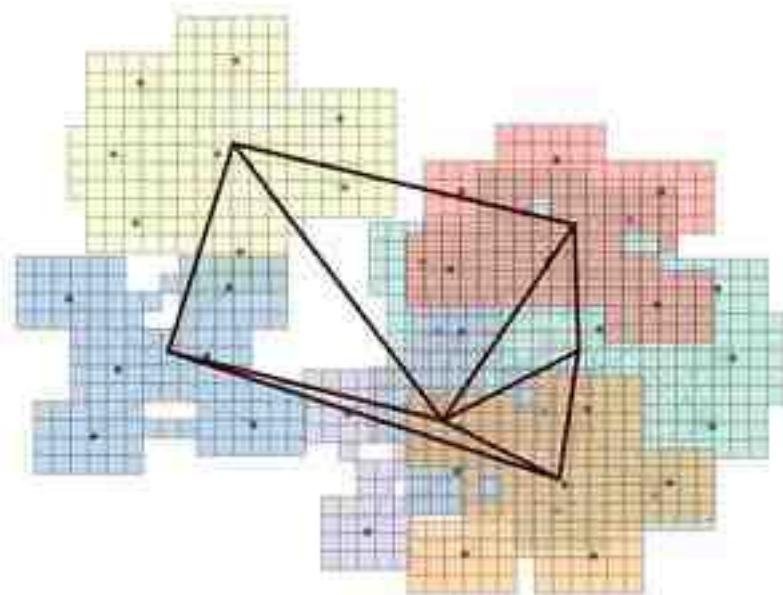


Stable
Packing alg.



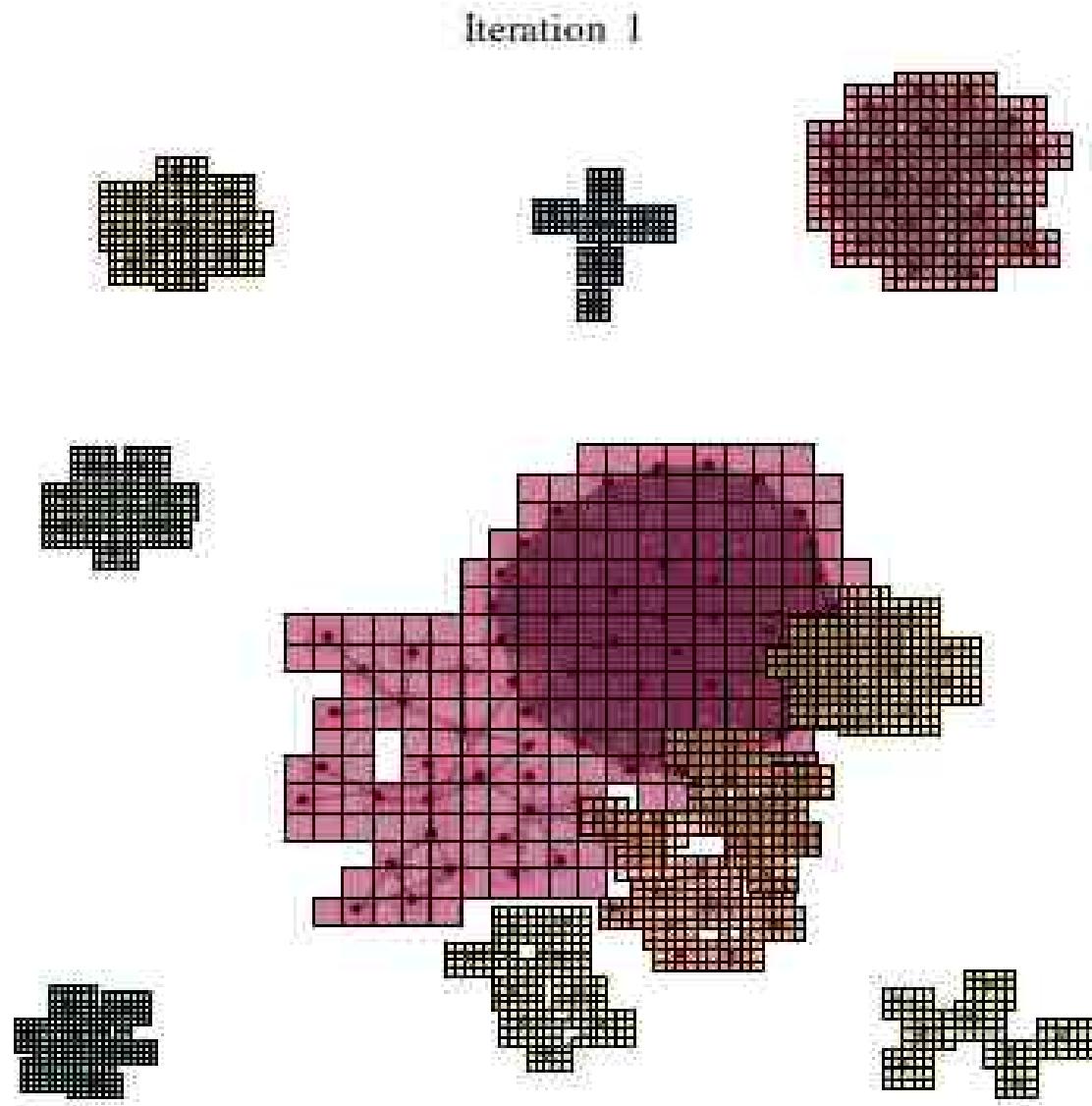
Stable Packing

- Use “scaffold” to maintain the relative positions



Stable Packing

- Animate over 10 iterations



TwitterScope

- The algorithms are applied to an online application – TwitterScope
- Monitor keywords
- Push to the browser in a streaming fashion
- ~300 tweets at a time
- For keywords like “news”, most of the tweets and refreshed. Stability is impossible.
- For keywords like “visualization”, only a few new tweets per minutes – stability comes into play

Conclusion

- Significant progress in algorithms for drawing large graphs in the last 10 years
- Challenges remain due to ever increasing size and complexity of graphs
- Making visualization in familiar metaphor can make complex data accessible to a larger audience (e.g., the Map of Music recorded 640K hits on stumbleupon.com)