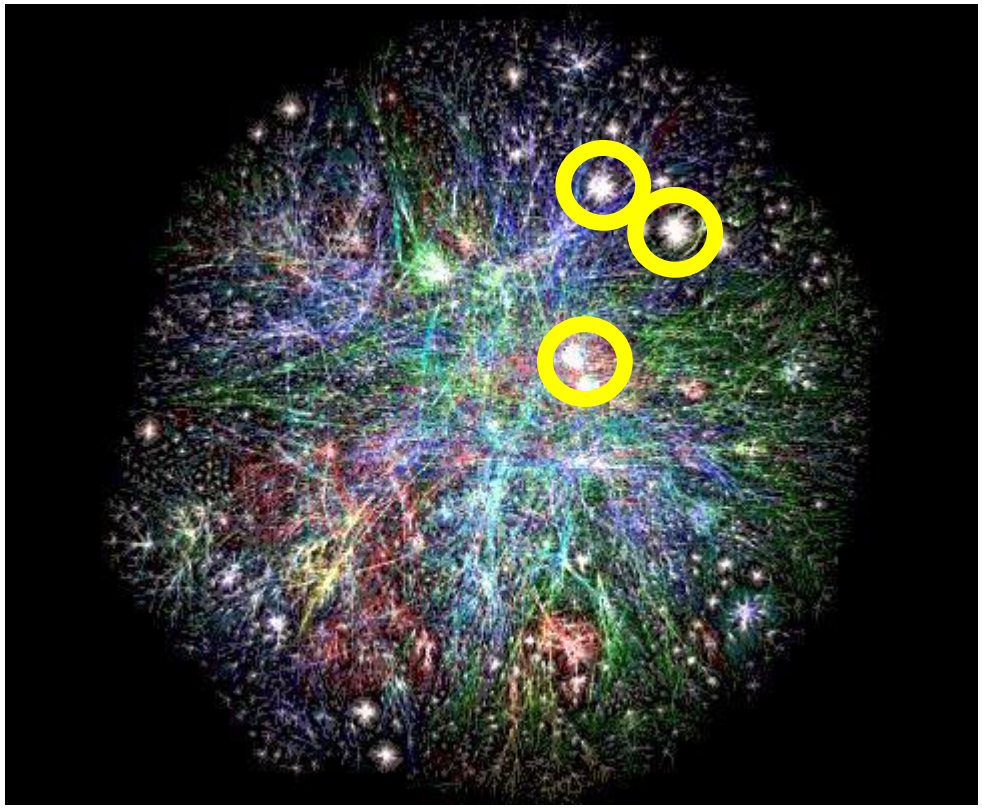
Algorithms for Large Graphs

Marc Bux

Web Graph



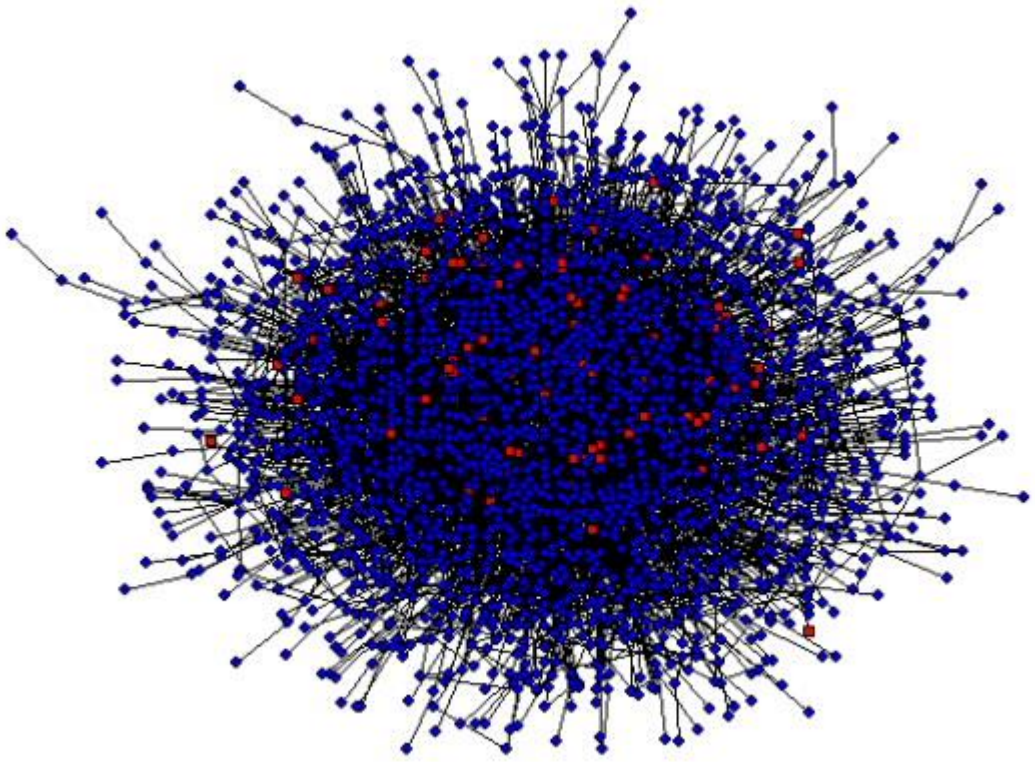
Note the strong local clustering

This is not a random graph

- Graph layout is difficult

[http://img.webme.com/pic/c/chegga-hp/opte_org.jpg]

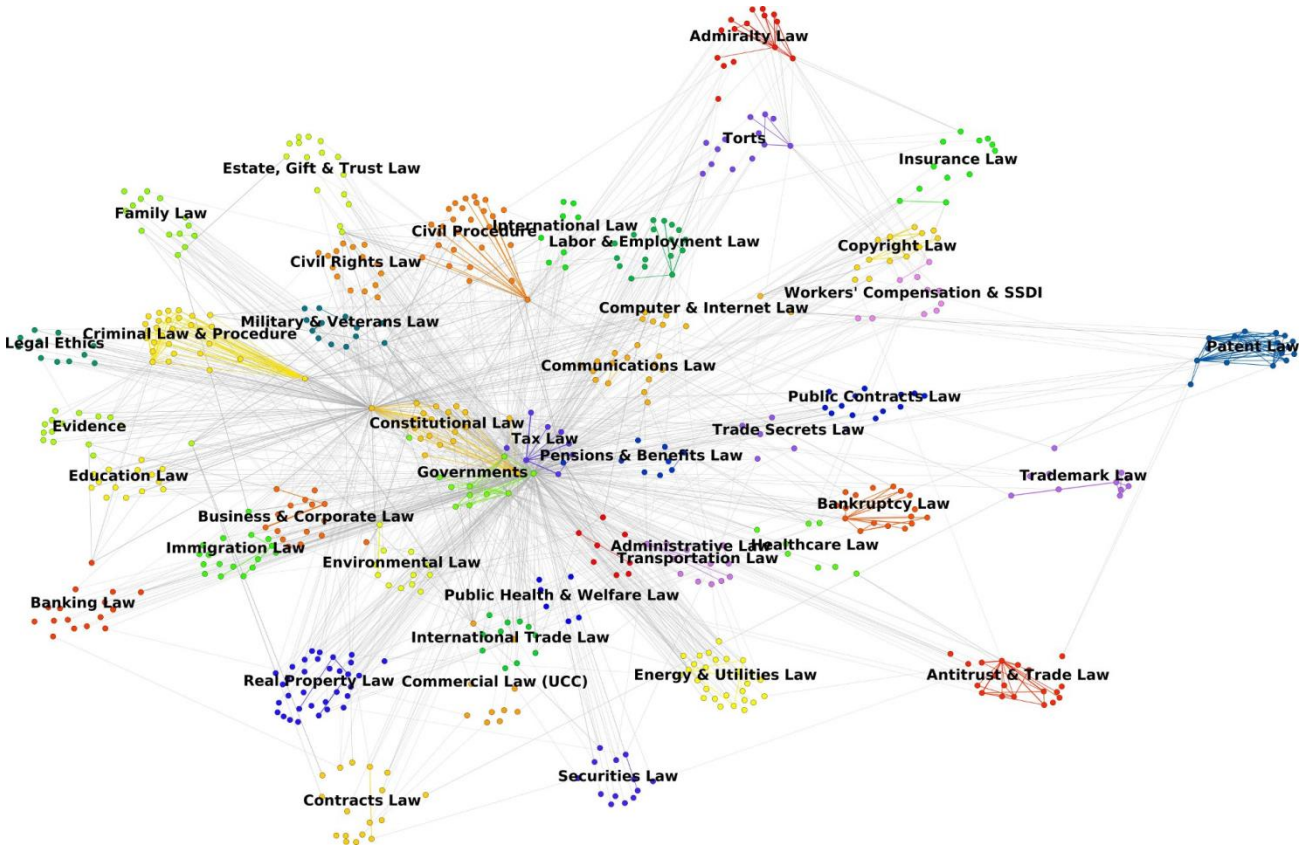
Human Protein-Protein-Interaction Network



- Still terribly incomplete
- Proteins that are close in the graph likely share function

[http://www.estradalab.org/research/index.html]

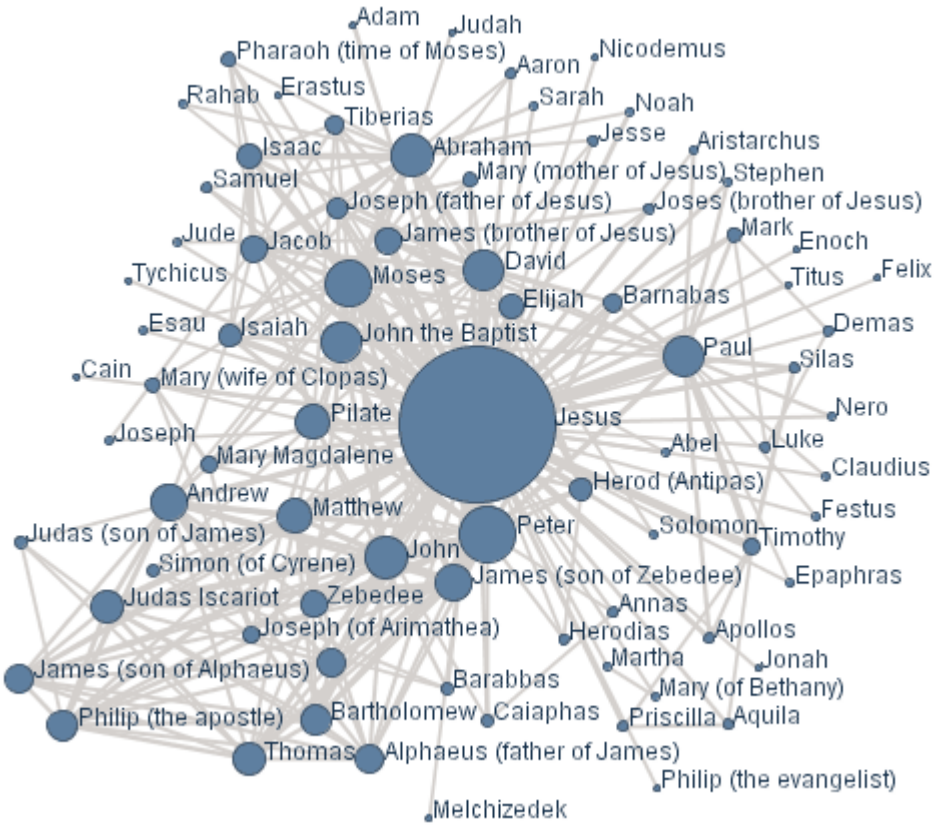
Word Co-Occurrence



- Words that are close have similar meaning
- Words cluster into topics

[http://www.michaelbommarito.com/blog/]

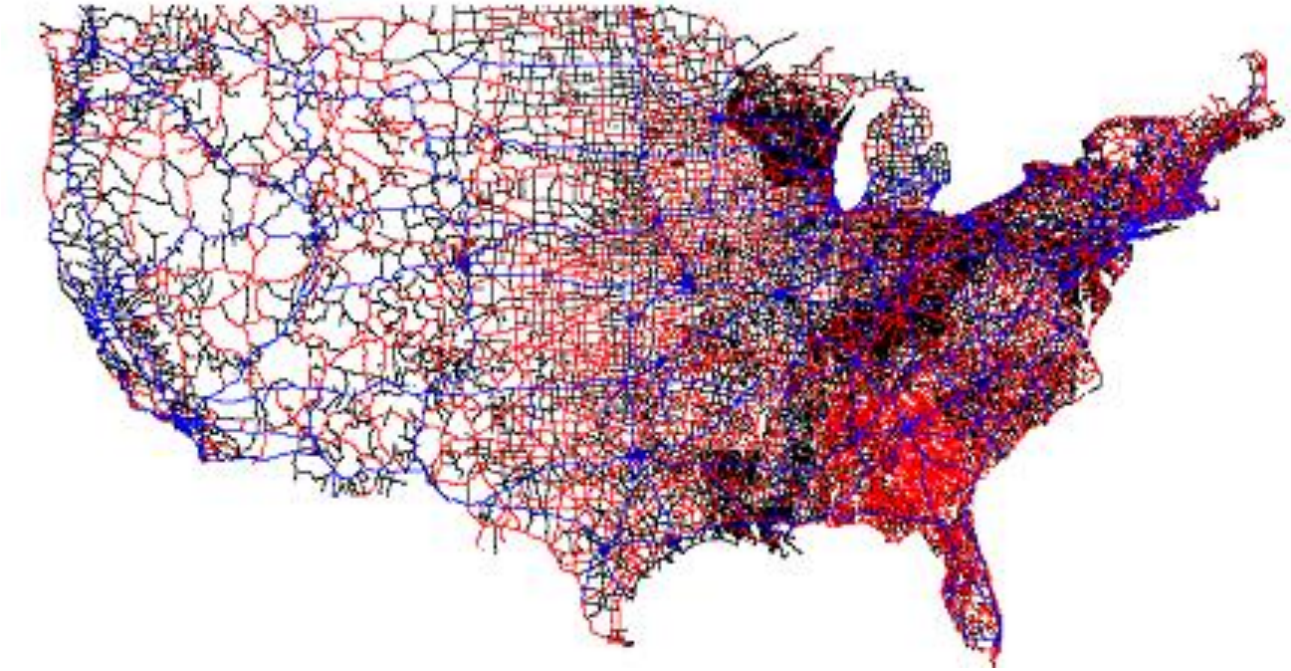
Social Networks



- Six degrees of separation

[http://tugll.tugraz.at/94426/files/-1/2461/2007.01.nt.social.network.png]

Road Network



- Specific property: Planar graphs

[Sanders, P. &Schultes, D. (2005).Highway Hierarchies Hasten Exact Shortest Path Queries. In *13th European Symposium on Algorithms (ESA), 568-579.*]

Graph Algorithms

- Have a long tradition

- Shortest paths (Route planning)
- Minimal spanning trees (network design)
- Connected components (Divide and conquer)
- Transitive Closure (reachability)
- Subgraph search (SPARQL, Semantic Web)
- Graph clustering, min-cut partitioning (cliques and groups)
- Graph centrality (speaker and hubs)
- …

Size Matters

- Classical research
  - Complexity of graph problems & algorithms
  - Graph in main memory – adjacency list or matrix
  - No graph preprocessing

- Here: Graphs are very large, tasks re-occur
  - Typical database scenario: Querying a graph
  - May not fit in memory: Reduce I/O, distribute data and/or query
  - Preprocessing pays off: Index structures
  - Modern hardware: Caches, NUMA

Some Large Graphs

cit-Patents	Directed, Temporal, Labeled	3,774,768	16,518,948	Citation network among US Patents
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
soc-LiveJournal1	Directed	4,847,571	68,993,773	LiveJournal online social network
com-Friendster	Undirected, Communities	65,608,366	1,806,067,135	Friendster online social network
email-Enron	Undirected	36,692	367,662	Email communication network from Enron
roadNet-CA	Undirected	1,965,206	5,533,214	Road network of California
wiki-meta	Edits	2.3M users, 3.5M pages	250M edits	Complete Wikipedia edit history (who edited what page)

Source: Stanford Large Network Dataset Collection

Who should be here

- Informatik
  - Master of Science
  - Master of Education
  - Diplom
- Ability to read English papers
- Knowledge in databases
  - Optimization, index structures, …
- Knowledge in algorithms
  - Trees, graphs, dynamic programming, complexity, …

How it will work

- Today: Presentation and selection of topics
- Obtain literature by 23.10.2015
  - Contact me if literature unavailable
- Make appointment with me by 2.11.2015 to meet me by 20.11.2015 to discuss topic and papers
  - One student a day, first-come-first-served
- Present topic in 5 min flash-presentation on 7.12.2015 at RUD 26, 0'313
- Make appointment with me by 21.12.2015 to meet me by 15.1.2016 to discuss slides
- Present your topic (30 / 45 min) at the Blockseminar on 29.1. and 5.2.2016
- Write seminar thesis (10 / 15 pages) by 31.3.2016

ToC

- Introduction
- Topics
- Assignment
- Hints on presenting your topic and writing your thesis

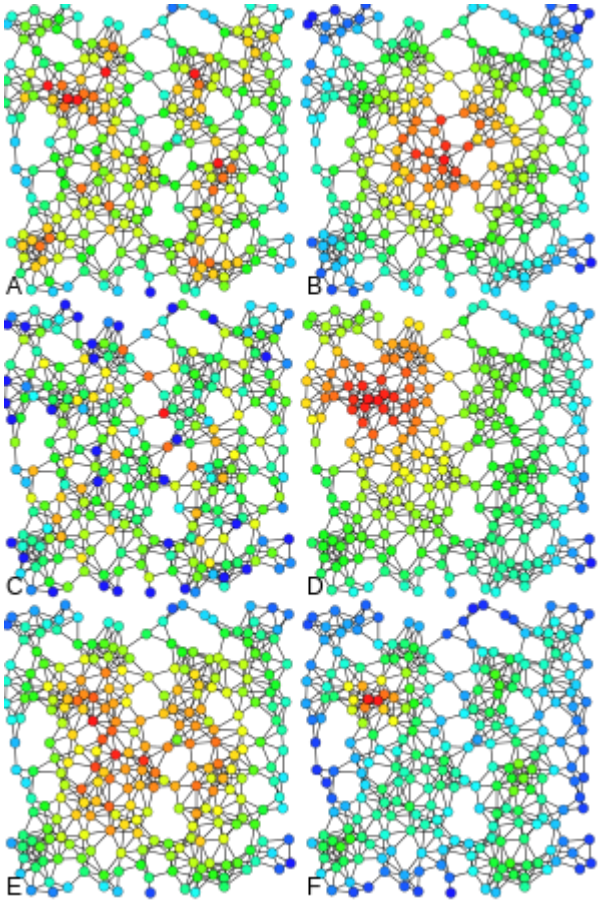
Introductory Literature

- Sedgewick, "Algorithms in Java", Part 5, Chapter 17

Betweenness Centrality	
Reachability	
Graph Algorithms on Pregel	
Balanced Graph Partitioning	
Graph Alignment	
Approximate Subgraph Search	
Regular Path Queries	
Ford-Fulkerson Algorithm	
Route Planning	

https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws1516/se_largegraphs

Betweenness Centrality

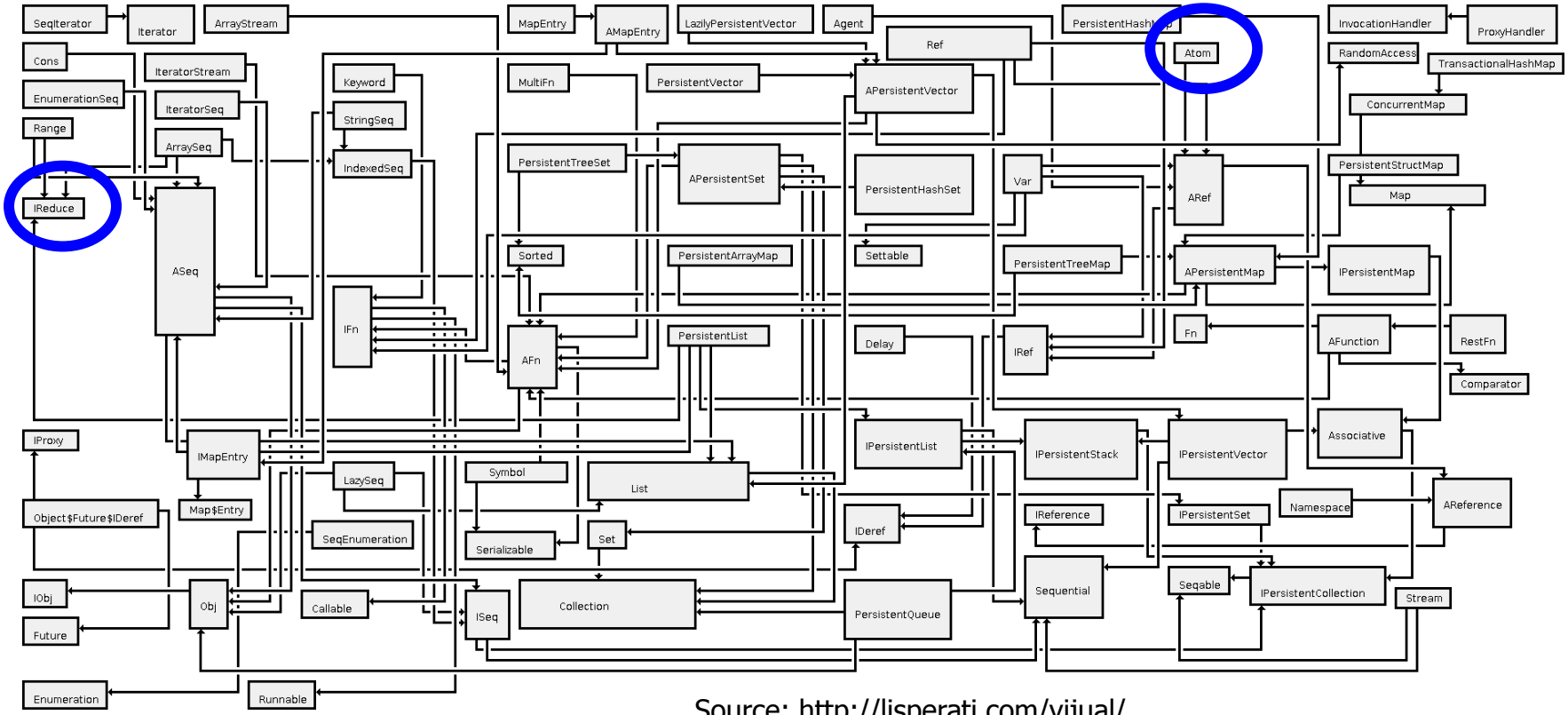


Source: Wikipedia

- Several definitions exist to capture „centrality" in graphs
- Betweenness Centrality: Number of shortest paths through a node

Examples of A) Degree centrality, B) Closeness centrality, C) Betweenness centrality, D) Eigenvector centrality, E) Katz centrality and F) Alpha centrality of the same graph.

Reachability

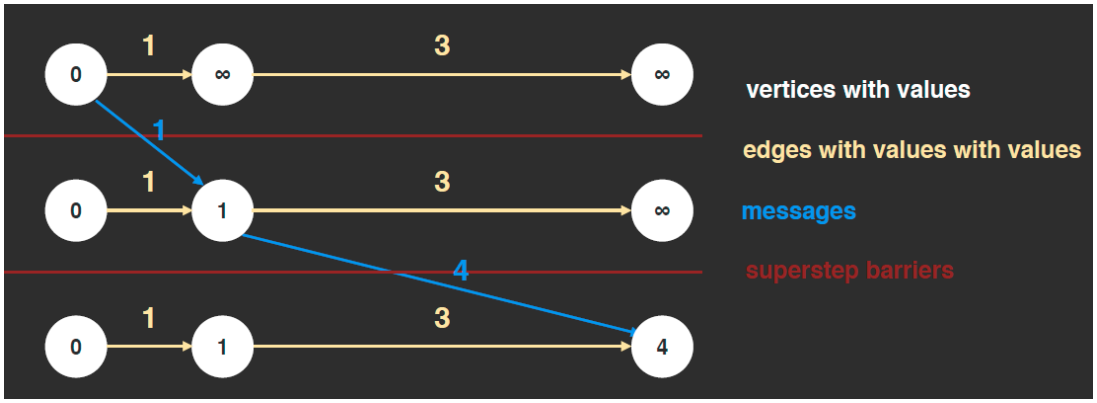


Source: http://lisperati.com/vijual/

- Does a path from a node to another exist?
- Label nodes such that labels carry reachability information
- Remember min-post order numbering

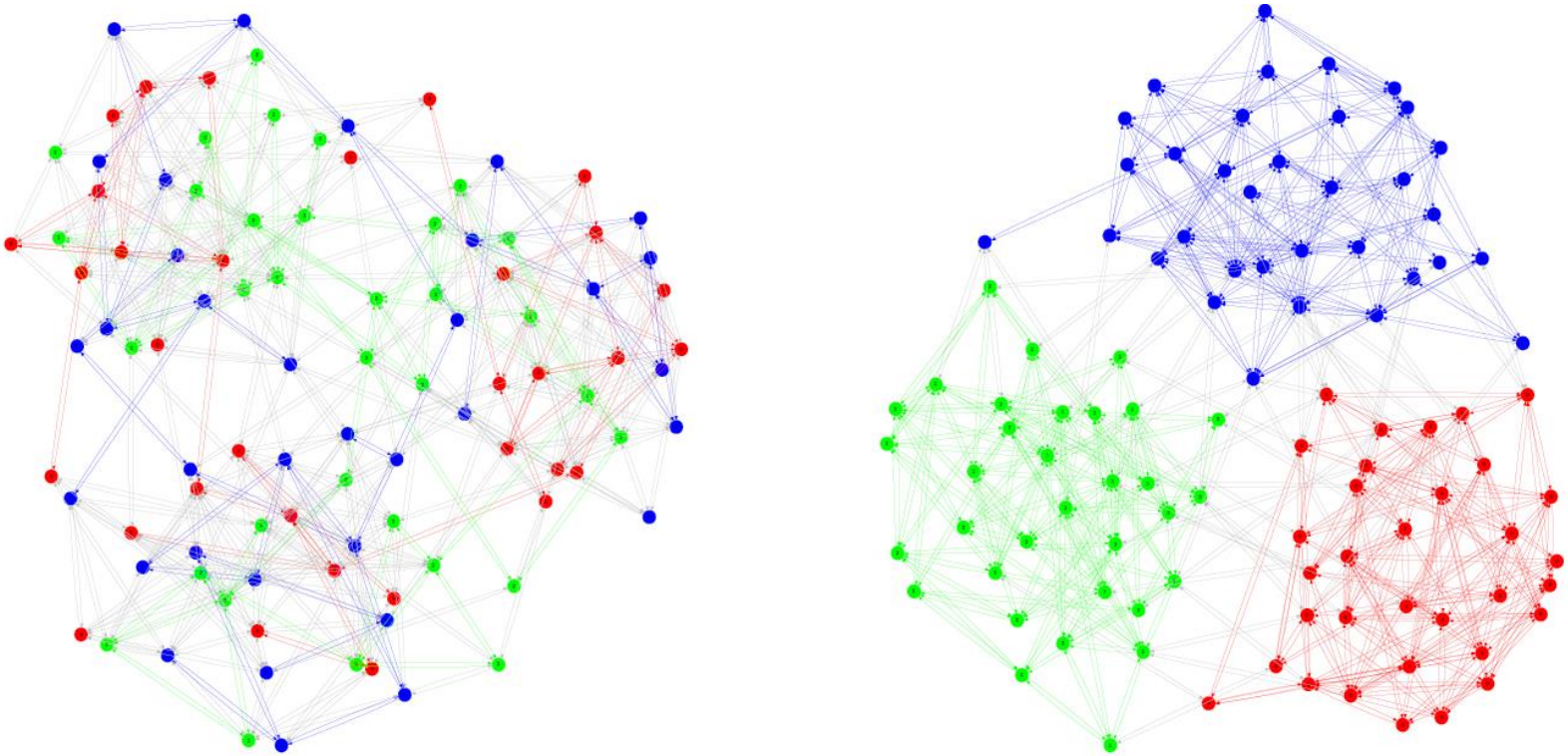
Graph Algorithms on Pregel

- Pregel / Giraph: „Think like a vertex“
- Vertices & edges store values
- Bulk synchronous parallel:
  - local computation (concurrent and independent on each vertex)
  - comunication (messages sent between vertices)
  - barrier synchronization
- Several algorithms have been adapted for Pregel / Giraph



Source: Mario Völker, Seminar Infrastrukturen für BIG DATA Anwendungen

Balanced Graph Partitioning

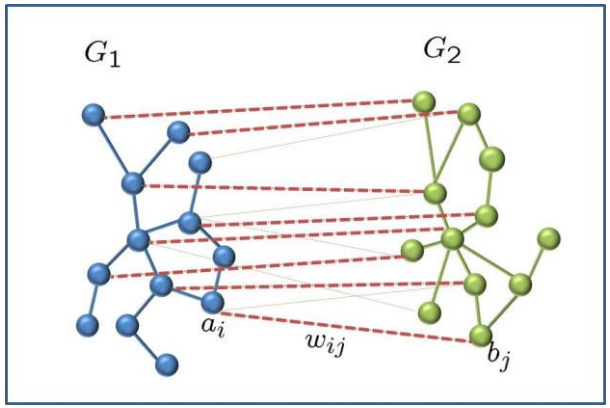


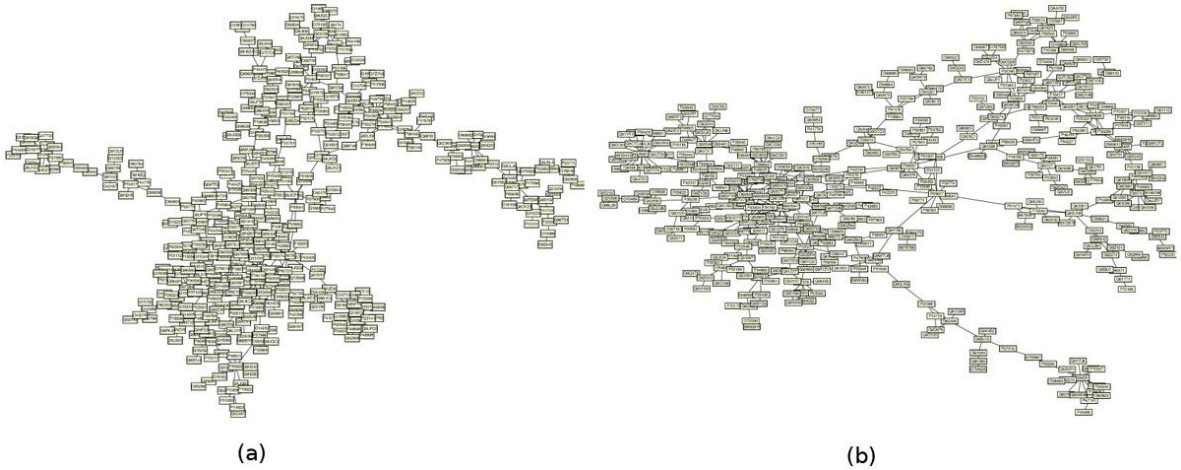
Source: Rahimian, Fatemeh, et al. "Ja-be-ja: A distributed algorithm for balanced graph partitioning." Self-Adaptive and Self-Organizing Systems (SASO), 2013 IEEE 7th International Conference on. IEEE, 2013.

- Divide graph into components with min. edges in between
- Components comprise equal number of nodes

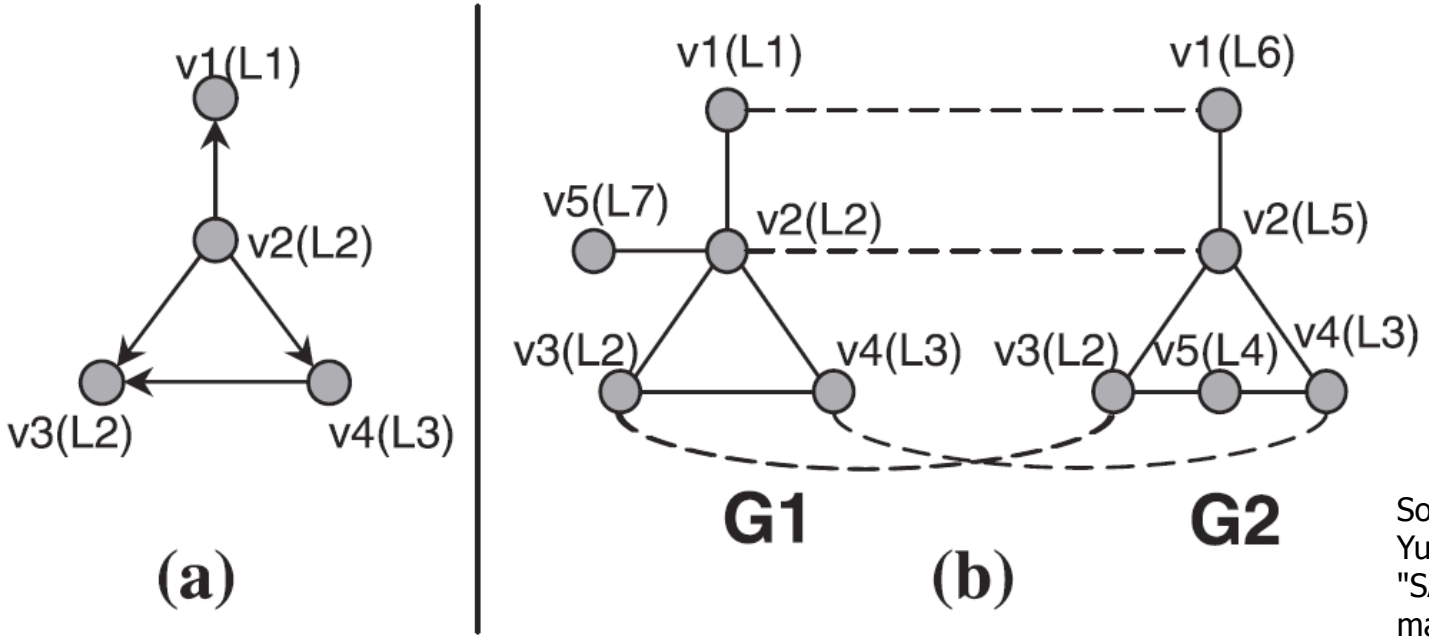
Graph Alignment

- Given two graphs G and H, find a mapping of nodes from g to nodes from h such that the topology is maximally preserved (intuitive definition)
- Highly complex problems, only approximations
- Heuristics, fast filtration





Approximate Subgraph Search



Source: Tian, Yuanyuan, et al. "SAGA: a subgraph matching tool for biological graphs." Bioinformatics 23.2 (2007): 232-239.

Fig. 1. (a) An example graph. (b) An example subgraph match.

- Find approximate matches of a query graph in set of graphs
- Penalize structural differences, node mismatches and gaps
- Index built for target set of graphs to speed up queries

Regular Path Queries

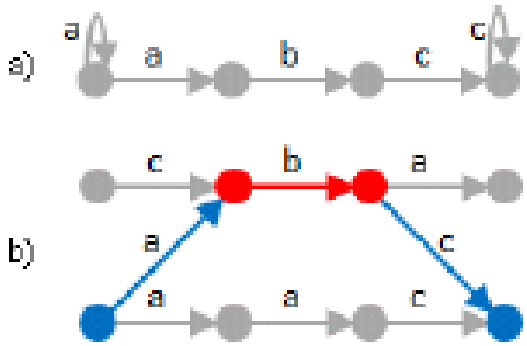


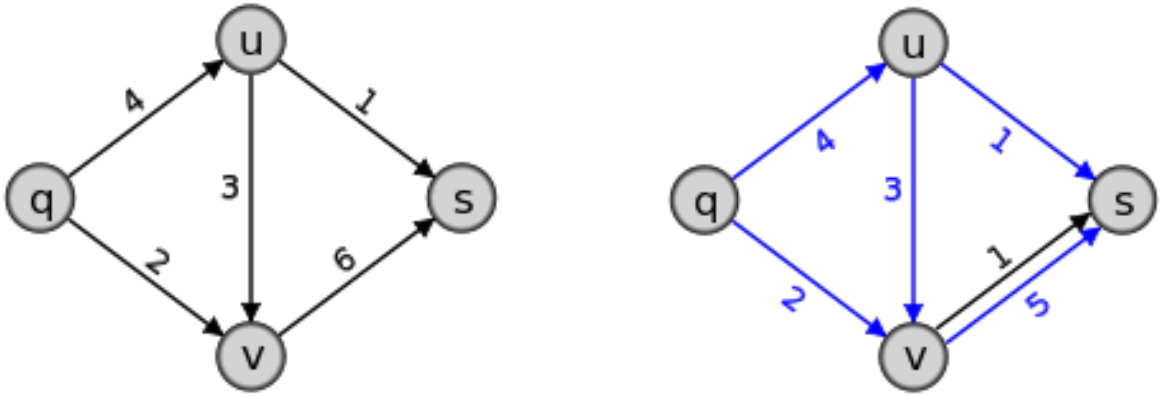
Figure 1: a) The RPQ $a+\ b\ c+$ shown as a (nondeterministic) automaton; b) a path fulfilling the RPQ in a small exemplary graph.

Source: Koschmieder, André, and Ulf Leser. "Regular path queries on large graphs." Proceedings of the 24th international conference on Scientific and Statistical Database Management (2012): 177-194.

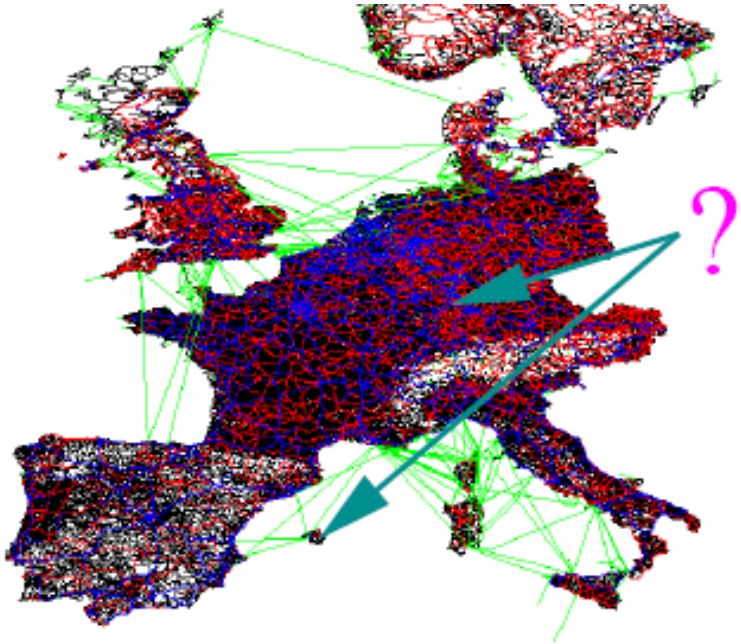
- RPQ: regular expression over a graph
- Evaluation (on arbitrary graphs) is NP-hard
- Automata work well for trees (XML), but not for graphs
- Core idea: split RPQ at rare graph labels

Ford-Fulkerson Algorithm

- Interpret edge labels as capacity / bandwidth (limit to throughput)
- Determine max-flow from a source to a sink
- Core idea: start with zero flow and slowly increase flow (along any path)



Route Planning



Source: Sanders, Schultes (2005). "Highway Hierarchies Hasten Exact Shortest Path Queries". 13th European Symposium on Algorithms (ESA): 568-579.

- Streets and highways: Planar graphs
- Exploit fast tracks
- Algorithm, complexity

Betweenness Centrality	
Reachability	
Graph Algorithms on Pregel	
Balanced Graph Partitioning	
Graph Alignment	
Approximate Subgraph Search	
Regular Path Queries	
Ford-Fulkerson Algorithm	
Route Planning	

https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws1516/se_largegraphs

ToC

- Introduction
- Topics
- Assignment
- Hints on presenting your topic and writing your thesis

Allgemeine Hinweise

- Dozenten sind ansprechbar!
  - Vorbesprechung des Themas
  - Folien durchgehen
  - Abgrenzung der Ausarbeitung
- Diskussion erwünscht
  - Keine Angst vor Fragen: Fragen sind keine Kritik
  - Eine Frage nicht beantworten können ist in Ordnung
- Tiefe, nicht Breite
  - Lieber das Thema einengen und dafür Details erklären
- Bezug nehmen
  - Vergleich zu anderen Arbeiten (im Seminar)

Allgemeine Hinweise

- Werten und bewerten
  - Keine Angst vor nicht ganz zutreffenden Aussagen – solange gute Gründe vorhanden sind
  - Begründen und argumentieren
  - Kritikloses Abschreiben ist fehl am Platz
- Literaturrecherche ist notwendig
  - Die ausgegebenen Arbeiten sind Anker
  - Weiterführende Arbeiten müssen herangezogen werden
  - Auch Grundlagen nachlesen
- Checkliste auf der Website beachten

Wie halte ich einen Seminarvortrag

1. Wenn man nun so einen Seminarvortrag halten muss, dann empfiehlt es sich, möglichst lange Sätze auf die Folien zu schreiben, damit die Zuhörer nach dem Vortrag aus den Folienkopien noch wissen, was man eigentlich gesagt hat.

2. Während so einem Vortrag schaut sowieso jeder zum Projektor, also kann man das selbst ruhig auch tun - damit kontrolliert man gleichzeitig auch, ob der Beamer wirklich alles projeziert, was auf dem Laptop zu sehen ist. Ausserdem kann man so den Strom für das Laptop-Display sparen.

3. Übersichtsfolien am Anfang sind langweilig, enthalten keinen Inhalt und nehmen den Zuhörern die ganze Spannung. Schliesslich gibt‘s im Kino am Anfang auch keine Inhaltsangabe.

4. Powerpoint kann viele lustige Effekte, hat tolle Designs und Animationen. Die sollte man zur Auflockerung des Vortrags unbedingt alle benutzen, um zu zeigen, wie gut man das Tool im Griff hat.

5. Nicht zu wenig auf die Folien schreiben. Man weiß ja nie, ob man sie nicht doch ausdrucken muss, und man kann so wertvolle Zeit sparen, wenn man nicht weiterschalten muss.

6. Man sollte versuchen, möglichst lange zu reden. Die Zeitvorgaben sind nur für die Leute, die nicht genug wissen - eigentlich will der Prüfer sehen, dass man sich auch darüber hinaus mit dem Thema beschäftigt hat.

Bloß keine Hervorhebungen im Text – sonst müssen die Zuhörer ja gar nicht mehr aufpassen!

Hinweise zum Vortrag

- ~30 Minuten inkl. Diskussion
  (45 Minuten für 2er-Gruppen)
- Klare Gliederung
- Ab und an Hinweise geben, wo man sich befindet
- Bilder und Grafiken; Beispiele
- Font: mind. 16pt
- Eher Stichwörter als lange Sätze
- Vorträge können auch unterhaltend sein
  - Gimmicks, Rhythmuswechsel, Einbeziehen der Zuhörer, etc.
- Adressat sind alle Teilnehmer, nicht nur die Betreuer
- Technik: Laptop? Powerpoint?

Hinweise zur Ausarbeitung

- ~10 Seiten (15 für 2er-Gruppen) bei normaler Schriftart und Zeilenabstand
- Eine gedruckte Version abgeben
  - Selbstständigkeitserklärung unterschreiben
- Eine elektronische Version schicken
- Referenzen: Alle verwendeten und nur die
  - Im Text referenzieren, Liste am Schluss
- Korrekt zitieren
  - Vorsicht vor Übernahme von kompletten Textpassagen; wenn, dann deutlich kennzeichnen
  - Aussagen mit Evidenz oder Verweis auf Literatur versehen
- Verwendung von gefundenen Arbeiten im Web
  - Möglich, aber VORSICHT

Hinweise zur Ausarbeitung (Fortsetzung)

- Gezielt und sachlich schreiben
  - Ausführungen zur „Philosophische Überlegungen zu Vorzügen probabilistischer Verfahren im Vergleich zu Dempster‘s Theory of Evidence“ oder zur „Anmerkungen zur Trivialisierung des politischen Diskurs für soziale Netzwerke unter besonderer Berücksichtigung von Twitter“ möglichst kurz halten
  - Füllwörter vermeiden (dabei, hierbei, dann, …)
  - Knappe Darlegung, präzise Sprache
- Eine gute Gliederung ist die halbe Miete
- Gerne auch auf Englisch
- Kommen Sie zu Aussagen
  - Vorteile, Nachteile, verwandte Arbeiten, mögliche Erweiterungen, Anwendbarkeit, eigene Erfahrungen, …

Format

- Benutzung unserer Latex-Vorlage (siehe Website)
- Nur eine Schriftart, wenig und konsistente Wechsel in Schriftgröße und –stärke
- Inhaltsverzeichnis
- Bilder: Nummerieren und darauf verweisen
- Referenzen:
  - [1] Yan, X., Yu, P. S. and Han, J. (2004). "Graph Indexing: A Frequent Structure-Based Approach". SIGMOD, Paris, France.
  - [YYH04] Yan, X., Yu, P. S. and Han, J. (2004). "Graph Indexing: A Frequent Structure-Based Approach". SIGMOD, Paris, France.
- Darf man Wikipedia zitieren?
  - Ja, aber nicht dauernd

Questions?

- Today: Presentation and selection of topics
- Obtain literature by 23.10.2015
  - Contact me if literature unavailable
- Make appointment with me by 2.11.2015 to meet me by 20.11.2015 to discuss topic and papers
  - One student a day, first-come-first-served
- Present topic in 5 min flash-presentation on 7.12.2015
- Make appointment with me by 21.12.2015 to meet me by 15.1.2016 to discuss slides
- Present your topic (30 / 45 min) at the Blockseminar (early or mid February)
- Write seminar thesis (10 / 15 pages) by 31.3.2016