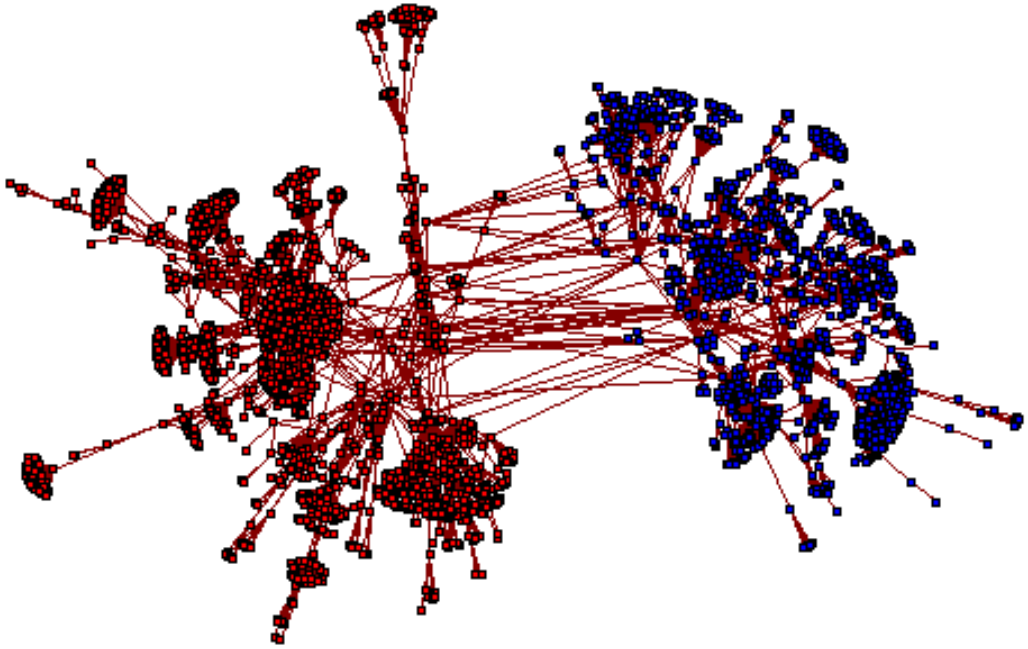
Finding and Visualizing Graph Clusters Using PageRank Optimization

Fan Chung and Alexander Tsiatas, UCSD

WAW 2010

What is graph clustering?

- The division of a graph into several partitions.
- Clusters should be selected according to some criteria:
  - For example: balance, graph connectivity and separation, …



Why use graph clustering?

- Identification of communities or logical subgroups within a larger network
  - Targeted advertising
  - Product recommendations

Customers Who Bought This Item Also Bought

Spectral Graph Theory (CBMS Regional Conferenc... by Fan R. K. Chung

Random Graph Dynamics (Cambridge Series in Statis... by Rick Durrett

Modern Graph Theory by Bela Bollobas

Why use graph clustering?

- Image segmentation and computer vision
  [Shi, Malik ‘00]
  - Identifying distinct objects or surfaces in an image
- Effective resistance to network epidemics
  [Chung, Horn, Tsiatas ‘09]
- Many applications in machine learning and data processing

Graph clustering algorithms

- *k*-means [MacQueen ’67; Lloyd ‘82]
  - NP-complete to solve exactly.
    - Many heuristic iterative algorithms
  - Requires a notion of pairwise distances.
    - Usually used for vector-based data.
    - Intractable to embed a graph into a low-dimensional space, or to cluster data in high-dimensional space.
    - Using the usual (shortest-path) graph distance as a metric is not discerning in real-world graphs with the **small-world phenomenon**: all distances are small!

Graph clustering algorithms

- Spectral clustering [Shi, Malik ‘00; Ng, Jordan, Weiss ‘02]
  - Relies on matrix computation, which can be intractable for large networks.
  - Splits graph into 2 parts
    - For more, recursively apply the algorithm.
    - We will develop an algorithm for *k* parts without recursive division.
- Markov clustering [Enright, Van Dongen, Ouzounis ‘02]
  - Also reliant on matrix computations.

Graph clustering algorithms

- Affinity propagation [Frey, Dueck ‘07]
  - A heuristic algorithm using pairwise distances.
- Local partitioning algorithms [Andersen, Chung, Lang ’06; Andersen, Chung ‘07]
  - Algorithms for finding one smaller cut within a network.
  - We will find a more balanced partition into *k* parts.
  - Does not require pairwise distances
- Many others… [Schaeffer ‘07]

Our contribution

- A new graph clustering algorithm:
  - Find *k* centers of mass using PageRank
    - Avoid the need for a high-dimensional embedding
  - Use centers to derive *k* clusters using Voronoi diagrams
  - Perform computations efficiently using inexpensive approximation algorithms
- A graph drawing algorithm:
  - Use PageRank to assist in determining node locations, highlighting local cluster structure
- PageRank helps overcome several problems!

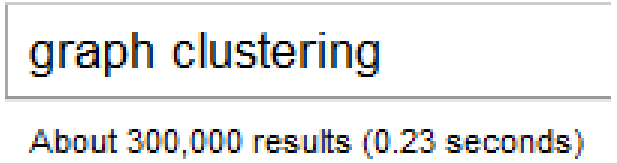
What is PageRank?

- Personalized PageRank [Brin, Page '98, Jeh, Widom '03] vectors quantify structural relationships between vertices and a specified starting distribution (or vertex) *s*:

$$\mathrm{pr}(\alpha, s) = \alpha s + (1 - \alpha)\mathrm{pr}(\alpha, s)W$$

- PageRank is the stationary distribution of a random walk (with transition probability matrix *W*) that restarts to *s* randomly.
  - Restart rate is controlled by the *jumping constant* α.

Why use PageRank?



- Proven to be effective and efficient at finding relevance in link-based data
- Intuitive interpretation of vertex relationships
  - The vth component of pr(α,u) quantifies how well-suited v is to be a representative center for u.
  - A natural metric for pairwise distances giving more information than simple graph distances.
  - Proven to be effective in Web search, local partitioning, combating epidemics….
- Performance
  - Using approximation algorithms [Andersen, Chung, Lang ’06; Chung, Zhao ‘10], PageRank vectors can be calculated efficiently.

Pairwise distances using PageRank

- In Euclidean space (for *k*-means):

$$\mathrm{dist}(u, v) = \|u - v\|_2$$

- Using PageRank:

$$\mathrm{dist}_\alpha(u, v) = \|\mathrm{pr}(\alpha, u)D^{-1/2} - \mathrm{pr}(\alpha, u)D^{-1/2}\|_2$$

where *D* is the diagonal degree matrix.

Throughout, we use node degrees and cluster volumes as normalizing factors.

- Generalizing to probability distributions *p,q* over vertices: $\mathrm{dist}_\alpha(p, q) = \sum_{u,v} p(u)q(v)\mathrm{dist}_\alpha(u, v)$

Centers of mass and clusters

- *c* is an *ε-center* or *center of mass* for a vertex set *S* if its PageRank distance to *S* is small:

$$\sum_{v \in S} \mathrm{dist}_\alpha(c, v) \leq \epsilon$$

  Here, *c* can be an individual vertex or a more general probability distribution.

- A set *C* of *k* centers determines a set of *k* clusters R_c for every *c* in *C*:

$$R_c = \{x \in V : \mathrm{dist}_\alpha(c, x) \leq \mathrm{dist}_\alpha(c', x) \text{ for all } c' \in C\}$$

  In other words, clusters are determined using a *Voronoi diagram* with PageRank distances and the centers *C*.

Evaluating centers and clusters

- We need some way to describe how "good" a set of centers *C* and their corresponding clusters R_c are.
- For *k*-means: $\mu(C) = \sum_{v \in V} \mathrm{dist}(v, c_v)^2$
- Using PageRank:

$$\begin{aligned} \mu(C) &= \sum_{v \in V} d_v \| \mathrm{pr}(\alpha, v) D^{-1/2} - \mathrm{pr}(\alpha, c_v) D^{-1/2} \|_2^2 \\ &= \sum_{v \in V} d_v \mathrm{dist}_\alpha(v, c_v)^2 \end{aligned}$$

- Here, c_v is the center of mass closest to *v*.

Evaluating centers and clusters

- $\mu(C)$ quantifies how well each center c in C represents its cluster R_c.
- We also need a metric for evaluating how well each cluster R_c is structurally different from the overall graph structure, using the random walk stationary distribution π.

$$\Psi_\alpha(C) = \sum_{c \in C} \mathrm{vol}(R_c) \mathrm{dist}_\alpha(c, \pi)^2$$

- If $\Psi_\alpha(C)$ is large, then the clusters are well-separated.

Evaluating a graph for clustered structure

- We interpret the PageRank vector $p = \mathrm{pr}(\alpha, v)$ for a vertex v to give the suitability of other vertices to be its center of mass. We define the *α-PageRank-variance*:

$$\Phi(\alpha) = \sum_{v \in V} d_v \mathrm{dist}_\alpha(v, \mathrm{pr}(\alpha, v))^2$$

- If $\Phi(\alpha)$ is small, then the PageRank vectors for v and p are close, indicating a clustered structure.

Evaluating a graph for clustered structure

- We also define the *α-cluster-variance*:

$$\Psi(\alpha) = \sum_{v \in V} d_v \mathrm{dist}_\alpha(\mathrm{pr}(\alpha, v), \pi)^2$$

- If $\Psi(\alpha)$ is large, then centers of mass predicted by PageRank vectors are far from the stationary distribution, also indicating a clustered structure.

Relationship between metrics

- The objective is to find a "good" set of centers *C:*
  - This means $\mu(C)$ is small and $\Psi_\alpha(C)$ is large.
- But if we use PageRank vectors to "guess" centers of mass, these metrics are similar to $\Phi(\alpha)$ and $\Psi(\alpha)$.
- If we take enough samples for centers of mass using PageRank, these metrics can be made arbitrarily close.

Relationship between metrics

$$\mu(C) = \sum_{v \in V} d_v \mathrm{dist}_\alpha(v, c_v)^2 \qquad \Psi_\alpha(C) = \sum_{c \in C} \mathrm{vol}(R_c) \mathrm{dist}_\alpha(c, \pi)^2$$

$$\Phi(\alpha) = \sum_{v \in V} d_v \mathrm{dist}_\alpha(v, \mathrm{pr}(\alpha, v))^2 \quad \Psi(\alpha) = \sum_{v \in V} d_v \mathrm{dist}_\alpha(\mathrm{pr}(\alpha, v), \pi)^2$$

- Thus, to find small $\mu(C)$ and large $\Psi_\alpha(C)$, we must find an α that gives small $\Phi(\alpha)$ and large $\Psi(\alpha)$.

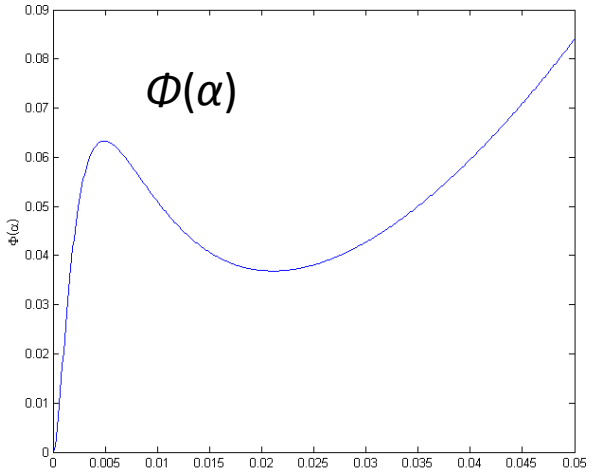
Selecting the jumping constant α

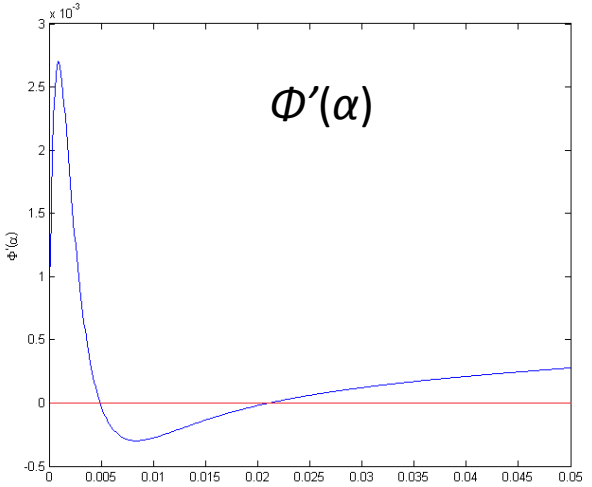
- We need to choose α so that $\Phi(\alpha)$ is small.

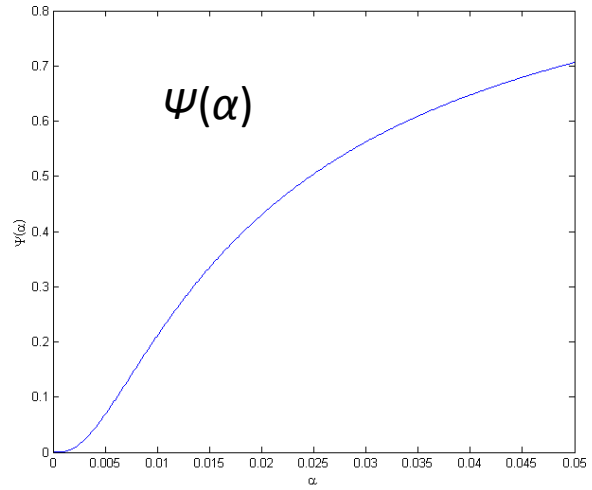
  (We will see later that if a graph has a clustered structure, $\Psi(\alpha)$ will be large.)
- We can find local minima of $\Phi(\alpha)$ by finding roots of $\Phi'(\alpha)$.
- We can also just use binary search over (0,1).
- There can be several good values for α, indicating a layered clustering structure.

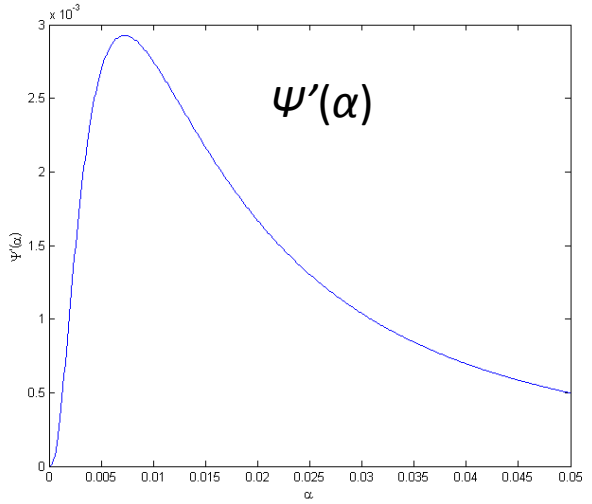
Selecting the jumping constant α: an illustration

- Let G be a dumbbell graph: two cliques of 20 nodes connected by a single edge.









The clustering algorithm

- **PageRank-ClusteringA**(G,k,ε):
  - For each vertex v, compute its PageRank vector pr(α,v)
  - For each root α of $\Phi'(\alpha)$:
    - If $\Phi(\alpha) \leq \varepsilon$ and $k \geq \Psi(\alpha) - 2 - \varepsilon$:
      - Select c log n sets of k potential centers, randomly selected from π. (Here, c is some large constant.)
      - For each set $S = \{v_1, \ldots, v_k\}$, let C be the set of centers of mass where $c_i = \text{pr}(\alpha, v_i)$.
      - If $|\mu(C) - \Phi(\alpha)| \leq \varepsilon$ and $|\Psi_\alpha(C) - \Psi(\alpha)| \leq \varepsilon$, return the clusters given by the k Voronoi regions according to the PageRank distances using C.
      - Otherwise, there may be no output – the graph does not have a clustered structure

Analysis of the clustering algorithm

- The algorithm **PageRank-ClusteringA** does not always return a clustering, but we will show that it does for a special class of **(*k*,*h*,β,*ε*)-clusterable** graphs *G*:
  - *G* can be partitioned into *k* parts so that each part *S* satisfies:
    - *S* has Cheeger ratio at most *h*.
    - *S* has volume at least β vol(*G*)/*k*.
    - There is a subset *S'* with vol(*S'*) ≤ (1-ε) vol(*S*) and Cheeger ratio at least $\sqrt{h/\log n}$.
  - Here, the Cheeger ratio for a set *H* is the ratio of the number of edges leaving *H* and vol(*H*).

Analysis of the clustering algorithm

- **Theorem.** Suppose a graph G has a (k,h,β,ε)-clustering and α,ε in (0,1) satisfy $\varepsilon \geq hk/2\alpha\beta$. Then with high probability, **PageRank-ClusteringA** returns a set C of k centers with $\Phi(\alpha) \leq \varepsilon$, $\Psi(C) > k - 2 - \varepsilon$, and the k clusters are near optimal according to $\mu(C)$ with an additive error term ε.

Analysis of the clustering algorithm

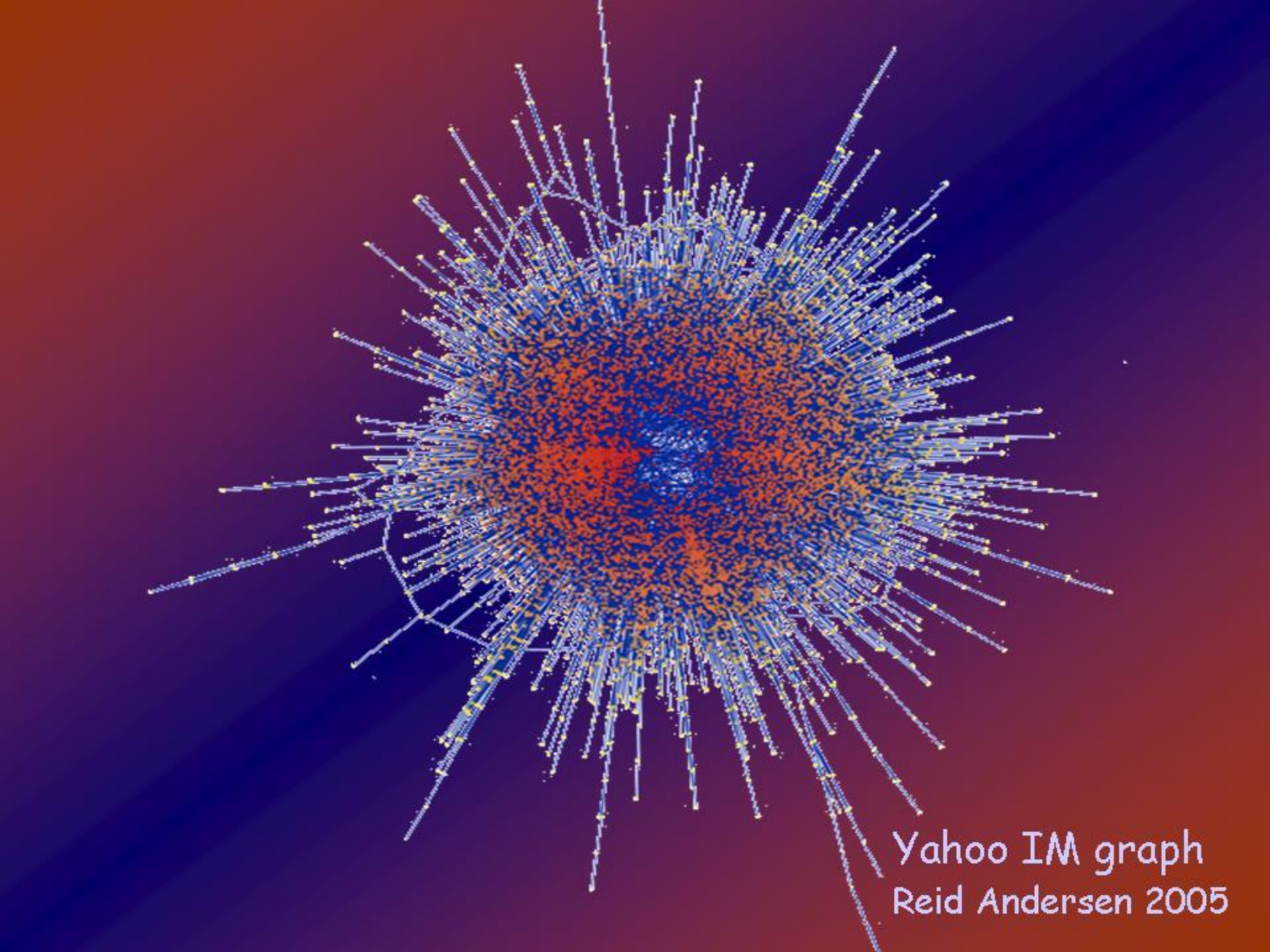
- The theorem mainly follows from the definition of (k,h,β,ε)-clustering, the discussed theory surrounding the evaluative cluster metrics Φ and Ψ, and probabilistic sampling arguments.
- The rest follows from the following claim:
  - If G can be partitioned into k clusters with Cheeger ratio at most h and $\varepsilon \geq hk/2\alpha\beta$, then $\Psi(a) \geq k - 2 - \varepsilon$.
  - This claim can be proven using a generalization of a known connection between PageRank and the Cheeger ratio [Andersen, Chung, Lang '06].

Analysis of the clustering algorithm

- The computational complexity of **PageRank-ClusteringA** is dominated by several computations:
  - Finding the roots of $\Phi'(\alpha)$
  - $O(k \log n)$ calculations of $\mu(C)$ and $\Psi_\alpha(C)$
  - $O(n)$ PageRank vector calculations
- These computations can be expensive, but fortunately we have inexpensive approximation algorithms:
  - Finding roots and calculating functions using sampling techniques [Rudelson, Vershynin '07]
  - Using approximate PageRank vectors [Andersen, Chung, Lang '06; Chung, Zhao '10]
- These techniques are outlined in an algorithm **PageRank-ClusteringB**.

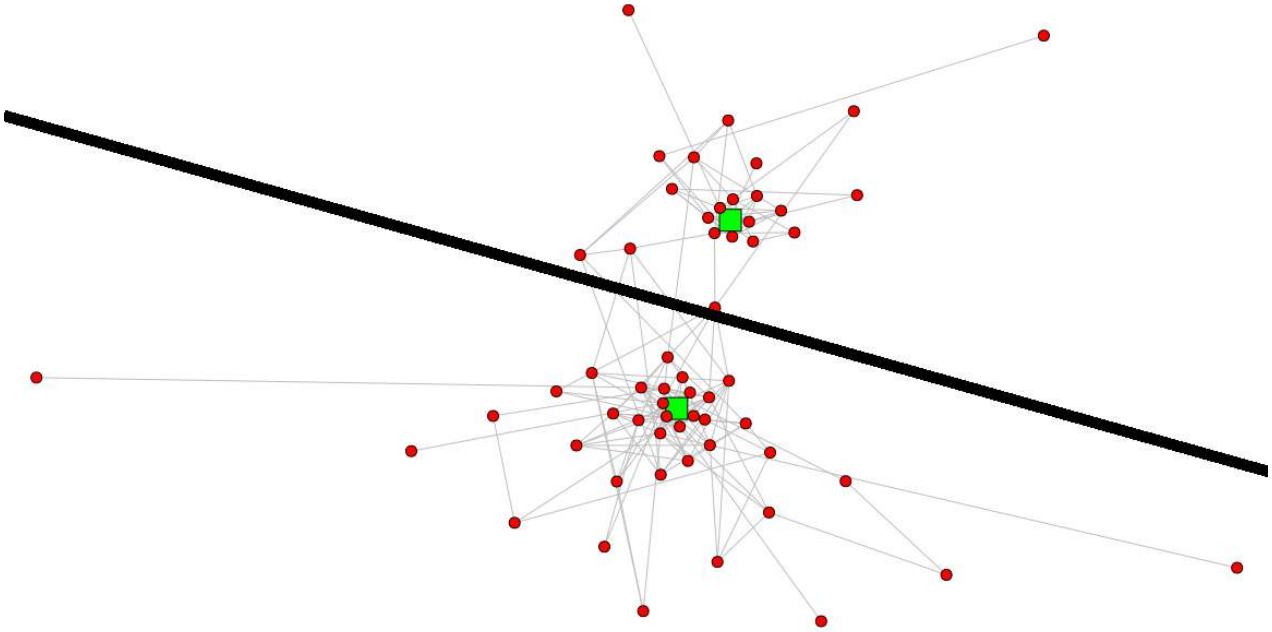
PageRank and graph visualization

- Many graph visualization algorithms have trouble showing local structure in complex networks without resorting to hierarchical layouts.
- PageRank's quantitative information can be used to assist force-based graph layout algorithms [Kamada, Kawai '89], highlighting local clusters around *k* centers of mass.
  - For each center *c* and every non-center *v*, simulate a spring with force inversely proportional to the *v*th component of pr(α,*c*).
  - For two centers *c* and *c'*, simulate a spring with a strong repellent force.
  - We also overlay a Voronoi diagram in Euclidean space to highlight the clusters.


Yahoo IM graph
Reid Andersen 2005

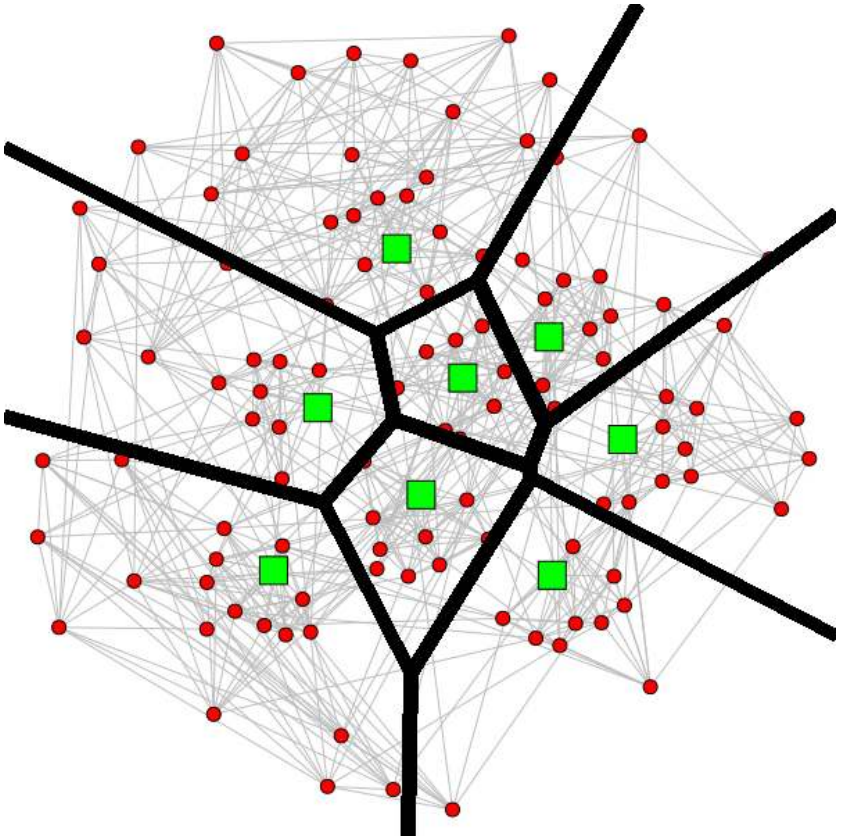
PageRank and graph visualization: an example

- Social network of dolphins [Newman, Girvan ‘04] with 2 clusters



PageRank and graph drawing: example

- Network of NCAA Division I football opponents [Girvan, Newman ‘02], highlighting several conferences



Open questions

- Improved performance and scalability
  - Graph visualization bottlenecks
- Applications of the graph clustering algorithm in specific settings
  - Biological graphs
  - Social networks

Thank you!

- Questions?