Diffusion and Clustering on Large Graphs

Alexander Tsiatas
Final Defense
17 May 2012
Introduction

Graphs are omnipresent in the real world – both natural and man-made

Examples of large graphs:

– The World Wide Web
– Co-authorship networks
– Telephone networks
– Airline route maps
– Protein interaction networks
Introduction

Real world graphs are **large**
- Internet: billions of webpages
- US Patent citations: over 3 million patents
- Millions of road miles in the USA

Computations become intractable
Important to have rigorous analysis
Defense outline

Background on personalized PageRank
Background on clustering
The PageRank-Clustering algorithm:
  An intuitive view
  More detailed analysis
Performance improvements
Examples
Random walks

Model of diffusion on $G$

- Start at $u$
- Move to $v$ chosen from neighbors
- Can repeat for any number of steps

PageRank is based on random walks
Random walk stationary distribution

As time goes to infinity:
- Constant probability on each node
- Proportional to degree distribution

Probability that random walk is at $u$?
Another model for diffusion on $G$

At each time step:

- Probability $(1 - \alpha)$: take a random walk step
- Probability $\alpha$: restart to distribution $s$
Personalized PageRank

A vector with $n$ components
  Each component: a vertex $u$

2 parameters:
  – Jumping constant $\alpha$
  – Starting distribution $s$
    Can be a single vertex

Denoted by $\text{pr}(\alpha, s)$
Personalized PageRank distributions

$\alpha = 0.1$  

$\alpha = 0.01$
Personalized PageRank distributions

$\alpha = 0.1$  $\alpha = 0.01$
Computing PageRank vectors

**Method 1:** Solve matrix equation:
\[
pr(\alpha, s) = \alpha s + (1 - \alpha) pr(\alpha, s) W
\]
- \( W \) is the *random walk matrix*
- Intractable for large graphs

**Method 2:** Iterate diffusion model
- Fast convergence, but still intractable for large graphs

**Method 3:** Use \( \varepsilon \)-approximate PageRank vector
- Uses only local computations
- Running time: \( O(1/\varepsilon\alpha) \) independent of \( n \)
- [ACL06; CZ10]
Defense outline

Background on personalized PageRank

Background on clustering

The PageRank-Clustering algorithm:
   An intuitive view
   More detailed analysis
   Performance improvements
   Examples
Graph clustering

Dividing a graph into clusters

- Well-connected internally
- Well-separated
Applications of clustering

Finding communities
Exploratory data analysis
Image segmentation
etc.
Global graph partitioning

Given a graph $G = (V, E)$, find $k$ centers and $k$ corresponding clusters

Want two properties:

– Well-connected internally

– Well-separated
Diffusion, clustering, and bottlenecks

Hard to diffuse through bottlenecks
Bottlenecks are good cluster boundaries
Graph clustering: algorithms

**k-means** [MacQueen67; Lloyd82]

Attempts to find $k$ centers that optimize a sum of squared Euclidean distance

Optimization problem is NP-hard

Many heuristic algorithms used

Can get stuck in local minima
Graph clustering: algorithms

**Spectral clustering** [SM00; NJW02]

Requires an expensive matrix computation
Uses $k$-means as a subroutine
(in a lower dimension)
Graph clustering: algorithms

Local partitioning algorithms

Based on analyzing random walks [ST04] or PageRank distributions [ACL06]

Efficiently find (w.h.p.) a local cluster around a given vertex

Can be stitched together to form a global partition – more expensive
Defense outline

Background on personalized PageRank
Background on clustering

The PageRank-Clustering algorithm:
  An intuitive view
  More detailed analysis
  Performance improvements
  Examples
Pairwise distances using PageRank

$k$-means requires *pairwise distances*

Graph setting: not Euclidean space

Graph distance (hops) is not very helpful

Real-world graphs have the **small-world phenomenon**

**PageRank distance:**

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

where $D$ is the diagonal degree matrix

Closely related vertices: close in PageRank distance
Using PageRank to find centers

**Observation**: components of $\text{pr}(\alpha, v)$ give a ranking of potential cluster centers for $v$
Using PageRank to find centers

**Observation**: components of $\text{pr}(\alpha, v)$ give a ranking of potential cluster centers for $v$
Evaluating clusters

We will develop 2 metrics for $k$ centers $C$:

$\mu(C)$ : measures internal cluster connectivity

$\Psi_\alpha(C)$ : measures separation

**Goal**: find clusters with small $\mu(C)$ and large $\Psi_\alpha(C)$
Internal cluster connectivity

\[ \mu(C') = \sum_{v \in V} d_v \text{dist}_\alpha(c_v, v)^2 \]

(similar to \(k\)-means)

\(c_v\) is the center closest to \(v\)

Small \(\mu(C)\): clusters are well-connected internally

Small distances to purple, large distances to orange

Small distances to orange, large distances to purple
Cluster separation

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

- \( R_c \) is the cluster containing \( c \)
- \( \pi \) is the stationary distribution for the random walk
- If \( \Psi_\alpha(C) \) is large, then the clusters are well-separated
Finding clusters using $\mu(C)$ and $\Psi_\alpha(C)$

Well-connected internally: $\mu(C)$ is small
Well-separated: $\Psi_\alpha(C)$ is large
Optimizing these metrics: computationally hard
New method to find clusters

Form $C$ randomly

**Step 1:** choose a set $C'$ of $k$ vertices sampled from $\pi$

**Step 2:** for each $v \in C'$, the *center of mass* $c_v$ is the probability distribution $\text{pr}(\alpha, v)$

Assign each vertex $u$ to closest center of mass using PageRank distance
Clusters should be well-connected internally

Before: want small $\mu(C)$

$$\mu(C) = \sum_{v \in V} d_v \text{dist}_\alpha(v, c_v)^2$$

Now: replace $c_v$ with a sample from $\Pr(\alpha, v)$.

Expectation of $\mu(C)$: (no dependence on $C$)

$$\Phi(\alpha) = \sum_{v \in V} d_v \text{dist}_\alpha(v, \Pr(\alpha, v))^2$$
Clusters should be well-separated

Before: want large $\Psi_\alpha(C)$

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

Now: replace $c$ with a sample from $\text{pr}(\alpha, v)$, for each vertex $v$

Expectation of $\Psi_\alpha(C)$: (no dependence on $C$)

$$\Psi(\alpha) = \sum_{v \in V} d_v \text{dist}_\alpha(\text{pr}(\alpha, v), \pi)^2$$
New metrics are tractably optimizable

Choose $C$ randomly and optimize expectations:

- Want small $\Phi(\alpha)$, large $\Psi(\alpha)$
- Only depends on $\alpha$
- No such $\alpha$? $G$ isn’t clusterable.
To find centers and clusters

**Step 1:** find an $\alpha$ with small $\Phi(\alpha)$, large $\Psi(\alpha)$

**Step 2:** randomly select $C$ using PageRank with this chosen $\alpha$

**Result:** $E[\mu(C)]$ is small, $E[\Psi_{\alpha}(C)]$ large

Choose enough sets $C$ so that metrics are close to expectation
The complete algorithm

**PageRank-Clustering**\((G, k, \varepsilon)\):

1. For each vertex \(v\), compute \(pr(\alpha, v)\)
2. For each root of \(\Phi'(\alpha)\):
   - If \(\Phi(\alpha) \leq \varepsilon\) and \(k \geq \Psi(\alpha) - 2 - \varepsilon\):
     1. Randomly select \(c \log n\) sets of \(k\) vertices from \(\pi\)
     2. For each set \(S = \{v_1, ..., v_k\}\), let \(c_i = pr(\alpha, v_i)\)
   3. If \(|\mu(C) - \Phi(\alpha)| \leq \varepsilon\) and \(|\psi_\alpha(C) - \Psi(\alpha)| \leq \varepsilon\) return \(C = \{c_1, ..., c_k\}\) and assign nodes to nearest center
3. If clusters haven’t been found, return nothing
Algorithm properties

**Correctness:** If $G$ has a clustered structure, algorithm finds clusters w.h.p.
Details in next section

**Termination:** Only finitely many roots of $\Phi'$
Finite amount of work for each root

**Running time:** $O(k \log n)$ computations of $\mu$ and $\psi_\alpha$; $O(n)$ PageRank computations
Defense outline

Background on personalized PageRank
Background on clustering
The PageRank-Clustering algorithm:
   An intuitive view
   More detailed analysis
Performance improvements
Examples
The previous section was not a proof

We can rigorously prove:

If $G$ has a clustered structure*,

1. There is an $\alpha$ such that $\Phi(\alpha)$ is small
2. There is an $\alpha$ such that $\Psi(\alpha)$ is large
3. With high probability, at least one potential $C$ has one vertex for each cluster
Definition: \(G\) is \((k, h, \beta, \epsilon)\)-clusterable if there are \(k\) clusters satisfying:

**Separation:** each cluster has Cheeger ratio \(\leq h\)

**Balance:** each cluster has volume \(\geq \beta \cdot \text{vol}(G) / k\)

**Connectivity:** for each subset \(T\) of a cluster \(S\):
if \(\text{vol}(T) \leq (1 - \epsilon) \cdot \text{vol}(S)\), then

Cheeger ratio \(\geq c(\beta, k, \epsilon) \cdot \sqrt{h \log n}\)

\(\text{vol}(S)\): sum of degrees of nodes in \(S\)

\(\text{Cheeger ratio}\): \(h_S = \frac{e(S, \bar{S})}{\min(\text{vol}(S), \text{vol}(\bar{S}))}\)
Definition: A PageRank vector computed from a restricted random-walk matrix \( W_\mathcal{S} \) rather than the full \( W \)

Notation: \( \text{pr}_\mathcal{S}(\alpha, \nu) \)
Vector with \( n \) dimensions – 1 for each vertex
\[ \text{pr}_\mathcal{S}(\alpha, \nu) = 0 \text{ for } \nu \text{ not in } \mathcal{S} \]

More about Dirichlet PageRank: in dissertation, not in this talk
\( G \) has clusters \( \Rightarrow \Phi(\alpha) \) is small

For the rest of this talk: assume \( \varepsilon \geq \frac{hk}{2\alpha\beta} \)

Interested in small \( h \), constant \( \beta \)

**Theorem:** if \( G \) is \((k, h, \beta, \varepsilon)\)-clusterable, \( \Phi(\alpha) \leq \varepsilon \)

Follows from a series of **Lemmas**
$G$ has clusters $\Rightarrow \Phi(\alpha)$ is small

**Theorem:** if $G$ is $(k, h, \beta, \varepsilon)$-clusterable, $\Phi(\alpha) \leq \varepsilon$

\[ \Phi(\alpha) = \sum_{v \in V} d_v \text{dist}_\alpha(v, \text{pr}(\alpha, v))^2 \]

**Overview:**

For each cluster $S$:

1. For most $v$ in $S$, $\text{pr}(\alpha, v)$ is concentrated in $S$
2. $\text{pr}_S(\alpha, v)$ and $\text{pr}_S(\alpha, \text{pr}_S(\alpha, v))$ are close
3. For most $v$ in $S$, $\text{pr}(\alpha, v)$ and $\text{pr}_S(\alpha, v)$ are close

1, 2, 3 lead to: $\text{pr}(\alpha, v)$ is close to $\text{pr}(\alpha, \text{pr}(\alpha, v))$  

$\Phi(\alpha)$: closeness in PageRank distance  

$\text{pr}(\alpha, v)$ gives best choices for $c_v$
For most $v$ in $S$, $\text{pr}(\alpha, v)$ is concentrated in $S$

**Lemma**: [generalization of ACL06]

There exists a $T \subseteq S$ with volume $\geq (1 - \delta) \text{vol}(S)$ with:

$$[\text{pr}(\alpha, v)](S) \geq 1 - \frac{h_S}{2\alpha\delta}$$

$S$: vertex set
$v$: vertex in $T$

**Separation property** $\Rightarrow$ RHS is large if $h$ is small
\( \text{pr}_S(\alpha, v) \) and \( \text{pr}_S(\alpha, \text{pr}_S(\alpha, v)) \) are close

**Lemma**: [generalization of ACL06 to Dirichlet PageRank]

For any \( v \) in \( S \) and integer \( t \geq 0 \):

\[
[\text{pr}_S(\alpha, v)](T) - [\text{pr}_S(\alpha, \text{pr}_S(\alpha, v))](T) \leq \alpha t + \sqrt{\text{vol}(T)} \left(1 - \frac{\varphi^2}{8}\right)^t
\]

\( S, T \): vertex sets with volume \( \leq \frac{1}{2} \text{vol}(G) \)

\( \varphi \): Cheeger ratio of a “segment subset”

(exact definition not needed)

**Connectivity** property \( \Rightarrow \varphi \) is large enough to make

RHS small
For most $v$ in $S$, $\text{pr}(\alpha, v)$ and $\text{pr}_S(\alpha, v)$ are close

**Lemma**: [from Chung10]

There exists a $T \subseteq S$ with volume $\geq (1 - \delta) \text{vol}(S)$ such that:

$$[\text{pr}(\alpha, v)](S') - [\text{pr}_S(\alpha, v)](S') \leq \sqrt{\frac{\epsilon'}{\delta}}$$

$S$: vertex set

$v$: vertex in $T$

$\epsilon' \geq c(\alpha) h_S$

**Separation** property $\Rightarrow \epsilon'$ is small if $h$ is small
\( G \) has clusters \( \Rightarrow \Phi(\alpha) \) is small

For each cluster \( S \):

1. For most \( v \) in \( S \), \( \Pr(\alpha, v) \) is concentrated in \( S \)
2. \( \Pr_S(\alpha, v) \) and \( \Pr(\alpha, \Pr_S(\alpha, v)) \) are close
3. For most \( v \) in \( S \), \( \Pr(\alpha, v) \) and \( \Pr_S(\alpha, v) \) are close

This implies:

For most \( v \) in \( S \), \( \Pr(\alpha, v) \) and \( \Pr(\alpha, \Pr(\alpha, v)) \) are:

1. Close for components in \( S \)
2. Concentrated in \( S \)
G has clusters $\Rightarrow \Phi(\alpha)$ is small

We have for each cluster $S$: 

For most $v$ in $S$, $\text{pr}(\alpha, v)$ and $\text{pr}(\alpha, \text{pr}(\alpha, v))$ are:

1. Close for vector components in $S$
2. Concentrated in $S$

$\text{pr}(\alpha, v)$ and $\text{pr}(\alpha, \text{pr}(\alpha, v))$ may be:

1. For all $v$: not close for vector components outside of $S$
2. For a few $v$ in $S$: not close at all
$G$ has clusters $\Rightarrow \Phi(\alpha)$ is small

Result of all this:

**Total variation distance** ($\Delta_{TV}$) between $\text{pr}(\alpha, v)$ and $\text{pr}(\alpha, \text{pr}(\alpha, v))$ is small

Related to **Chi-squared distance** ($\Delta_{\chi}$) [AF]:

$$\Delta_{TV} \leq \Delta_{\chi} \leq \sqrt{1 - (1 - 2\Delta_{TV})^2}$$

$\Delta_{TV}$ is small $\Rightarrow \Delta_{\chi}$ is small

$\Phi(\alpha) = \Delta_{\chi}^2$ so it’s small too!

(Can be shown to be $\leq \varepsilon$, assuming that constants are all chosen appropriately)
$G$ has clusters $\Rightarrow \Psi(\alpha)$ is large

**Theorem:** if $G$ is $(k, h, \beta, \varepsilon)$-clusterable, then $\Psi(\alpha) \geq k - 2 - \varepsilon$.

**Proof sketch:**

From previous **Lemma**: for most $v$, $pr(\alpha, v)$ is concentrated in $v$’s cluster

**Balance property $\Rightarrow \pi$ is not!**

$$
\Psi(\alpha) = \sum_{v \in V} d_v \text{dist}_\alpha(pr(\alpha, v), \pi)^2
$$
$G$ has clusters $\Rightarrow \psi(\alpha)$ is large
Randomly selecting centers works

Theorem:
Let $G$ be $(k, h, \beta, \varepsilon)$-clusterable
Choose $c \log n$ sets of $k$ vertices
A good set contains exactly 1 vertex from each cluster core
$\Pr[\text{At least one good set selected}] = 1 - o(1)$

Proof sketch:
Balance property $\Rightarrow$ each cluster $S$ has constant probability of being chosen
Previous Lemma $\Rightarrow$ the core of $S$ is at least a constant fraction of $S$ by volume
$\Pr[\text{Random set is good}]$ large enough so that $\Pr[\text{At least 1 out of } c \log n \text{ sets is good}] = 1 - o(1)$
Defense outline

Background on personalized PageRank
Background on clustering
The PageRank-Clustering algorithm:
  An intuitive view
  More detailed analysis
Performance improvements
Examples
PageRank-Clustering: Performance improvements

Use approximate PageRank algorithms to compute PageRank vectors [ACL06, CZ10]

An $\varepsilon$-approximate PageRank vector has error at most $\varepsilon \cdot \text{vol}(S)$ for any set of nodes $S$

Running time: $O(1/\varepsilon \alpha)$ independent of $n$
Defense outline

Background on personalized PageRank
Background on clustering

The PageRank-Clustering algorithm:
  An intuitive view
  More detailed analysis
  Performance improvements
  Examples
Clustering example

Social network of dolphins [NG04]

2 clusters
Clustering example

Network of Air Force flying partners [NMB04]
3 clusters
More complex example

- NCAA Division I football [GN02]
- Teams are organized into conferences
  - Drawing highlights several of them
Acknowledgements

Network epidemics
  Joint work with Fan Chung and Paul Horn
Global graph partitioning and visualization
  Joint work with Fan Chung
Trust-based ranking
  Joint work with Fan Chung and Wensong Xu
Planted partitions
  Joint work with Kamalika Chaudhuri and Fan Chung
Spectral analysis of communication networks
  Joint work with Iraj Saniee, Onuttom Narayan, and Matthew Andrews
Hypergraph coloring games and voter models
  Joint work with Fan Chung
Questions?