Charu C. Aggarwal
IBM T J Watson Research Center

Yan Xie, Philip S. Yu
University of Illinois at Chicago

Towards Community Detection in Locally Heterogeneous Networks

SIAM Conference on Data Mining, 2011

Introduction

- Community detection algorithms are used in a wide variety of social-networking applications
- Most social networks have varying levels of structural density in different parts of the network
- Global analysis can result in poor clustering when the heterogeneity in structural density is not accounted for
- In this paper we design a locally heterogeneous algorithm for community detection

Challenges of Local Heterogeneity

- The use of a uniform density criterion over the whole network will lead to imbalanced clusters
- Most nodes will be assigned to one or two super-clusters and the vast majority of communities may contain an insignificant number of nodes
- Established communities are often much more dense than emerging communities
- Since social networks are very large, it is challenging to perform local analysis on such a network

Local Community Detection: Goals

- The goal of local community detection is to design techniques which are sensitive to varying behavior of density in different parts of the network.

- Need to discover the relevant density at different parts of the network in parametric form and use it to discover important local regions

Contributions

- Design an effective algorithm for community detection in locally heterogeneous networks.

- Present the effects of local heterogeneity on community detection on real network data sets

- Illustrate advantages of local community detection over a global approach with case-studies

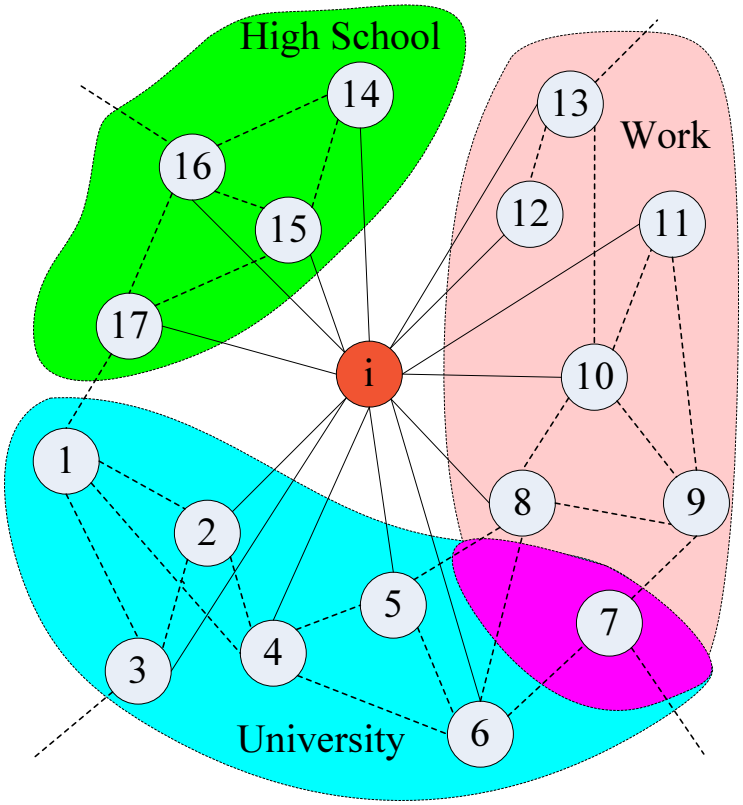
Some Observations

- Social networks are often quite sparse in terms of the number of links emanating from a particular node.

- The neighbors for a given node are also typically correlated with one another by linkage behavior.
  - Even in cases in which a node may have a large number of neighbors, these neighbors can be disjointed into a small number of correlated groups or communities.
  - The number of communities that a given node may belong to is usually quite small.

- We refer to this sparse and correlated property of social networks as the *local succinctness property*.

Illustration



- Illustration of local communities for users

Intuitions

- *Communities are typically formed as a result of the interaction between particular entities during specific periods of time.*
  - Different periods of time often lead to interactions with different geographical, interest, professional or student groups.
  - Often leads to communities which may have some overlap but are *largely* disjoint from one another.
- Friends within each of these categories tend have stronger ties with members of the same category, although cross-category ties are still possible.

Broad Approach

- Characterize the global structural behavior of the social network as a compact decomposition of the (succint) local behavior.

- Use a min-hash approach in order to determine a compact data structure representation which can be leveraged for finding a small number of local communities specific to each individual.

- The min-hash approach will exploit the local view of the social network for each node, and construct a *local projection* of that community for each individual.

- Merge the *local community projections* into a concise set of global communities.

Notations and Definitions

Symbol	Description
$A,\ B$	Node Sets.
$i,\ j$	Node indices.
$\mathcal{N}(i)$	Neighbor set of node i.
$\mathcal{N}(A)$	Neighbor set of node set A
$J(A,B)$	Pairwise Jaccard coefficient between A and B
$J(A_1 \dots A_n)$	Multi-way Jaccard similarity for $A_1 \dots A_n$
$JN(A,B)$	Pairwise Jaccard coefficient of neighbor sets, which is the same as $J(\mathcal{N}(A),\mathcal{N}(B))$
$JN(A_1 \dots A_n)$	Multi-way Jaccard similarity for neighbor sets of $A_1 \dots A_n$

Definitions

- Given a set of nodes $\mathcal{I} \equiv \{i_1, i_2, \ldots, i_K\}$ the group affinity is defined as the multi-way Jaccard similarity of **their neighbor sets**. This similarity is defined as follows:

$$JN(\mathcal{I}) := |\cap_{i \in \mathcal{I}} \mathcal{N}(i)| / |\cup_{i \in \mathcal{I}} \mathcal{N}(i)|.$$

- For a given node i, let $p_1(i) \ldots p_n(i)$ represent the pairwise group affinities for the n different 2-element sets containing i and each of the n different nodes.
  - The tail threshold $T(i)$ for node i is defined by $\mu(i) = \sum_{r=1}^{n} JN(\{i, r\})/n$.

Local Edge-based Communities

- A local edge-based community for node i is a **maximal** set of nodes $\mathcal{G}$ which satisfy the following conditions:
  - $\mathcal{G}$ contains node i
  - The edge-based group affinity is above the *tail threshold* $T(i)$ for node i.
  - In other words, we have $JN(\mathcal{G}) > T(i)$, and there is no superset $\mathcal{G}' \supset \mathcal{G}$ such that $JN(\mathcal{G}') > T(i)$.

Putting Together the Mosaic

- A compact set of communities is more useful for data mining purposes by consolidating related communities.
  - **Example:** In illustration, the local school-related community for any particular node may be pieced together in order to create a more coherent community.
- Once the local communities have been determined, we work with this set directly, and do not need to use edge-linkage behavior within this set.
- We use the set relationships between the different local communities in order to consolidate them.
- Determine patterns from the local communities which share a large number of nodes and perform the consolidation process.

Local Min-Hash Scheme for Community Detection

- Design a local min-hash scheme as a proxy for effective pattern sampling in a way which is sensitive to network locality.

- Technique has been used before for:
  - Frequent pattern mining
  - Global community detection

- Adapt approach for local community detection.

Local Community Detection

- We create a small-size description of the underlying communities in the data.
- Such an approach is flexible enough to accommodate both a local and global view.
- Provides a better understanding of how the communities relate both at the local and the global level.

Representation

- In order to represent the social network, we use the *conceptual representation of a* node-node adjacency matrix.

- For a network containing n nodes, we create a $n \times n$ matrix, in which the entry (i, j) is 1 if the node i is linked to node j. Otherwise the entry (i, j) is set to 0.

- All diagonal entries are always set to 1.

- Since we assume that "friendship linkages" are bi-directional, this matrix is symmetric in nature.

- In practice, this representation cannot be used efficiently, because the matrix is very sparse.

Min-Hash Approach: Broad Idea

- We sort the rows of this adjacency matrix, and determine the index of the first row for each column for which *any of these entries* are 1.

- It can be shown that the probability that these indices are the same for a pair of columns i and j is equal to the Jaccard coefficient used to measure the affinity between two nodes.
  - The denominator of the Jaccard Coefficient corresponds to a union event on set membership.
  - The numerator corresponds to the intersection event.

- The intersection event occurs if and only if all the min-hash indices for that set of columns are the same.

Min-Hash Approach

- It is possible to estimate the Jaccard coefficient by computing the fraction of this event occurrence over k samples.
- In practice, the actual matrix is not used, but we can work with only the edges incident to a node.
- We construct k different random sort-orders of the nodes.
- For each node i and the pth sort-order ($p \in \{1 \ldots k\}$), we examine its links, and determine the node index $Q(p, i)$ for the first node *linked* to i in this sort order.
- Thus, for each node i, we determine k different minimum indices, which are denoted by $Q(1, i)$, $Q(2, i)$ $\ldots$ $Q(k, i)$. $\Rightarrow$ This creates a matrix $\mathcal{M}$ of size $k \times n$.

Observations

- For a given set $S = \{i_1 \ldots i_r\}$, the Jaccard-coefficient $\mathcal{AJ}(S)$ for the set is given by the fraction of rows from $\mathcal{M}$, such that each such row j satisfies the following relationship:

$$Q(j, i_1) = Q(j, i_2) = \ldots = Q(j, i_r)$$

- The min-hash index is used in order to construct a *transactional representation* of the underlying data.

- *For each row*, we partition the set $Q(j, 1) \ldots Q(j, n)$ into groups for which the min-hash index values are the same.

- We can construct transactions $T_1 \ldots T_h$ corresponding to the different equi-index partitions from a single row in order to create new data base $\mathcal{T}$.

Transactional Representation

- Let $\mathcal{T}$ be the transactions constructed from the min-hash index set.

- The group affinity of a set of nodes S is equal to the absolute support of S in $\mathcal{T}$ divided by k.

Determining Local Communities

- The local communities are defined based on a local threshold $T(i)$.

- This threshold $T(i)$ translates into an *item-specific* (or more accurately *node-specific*) support for the frequent pattern mining problem.

- Determine any locally frequent pattern P from transaction set $\mathcal{T}$ with respect to local supports $T(1), \ldots, T(n)$, such that the support of P in $\mathcal{T}$ is at least $\min_{i \in P} T(i)$.

Local Communities with Pattern Mining

- There are tremendous numbers of overlaps in the local frequent patterns for different nodes, especially if the values of $T(i)$ for different nodes are close together.

- Not efficient to use frequent pattern mining for individual thresholds. A better solution is to *consolidate* the determination of frequent patterns.

- Approach proposed in Liu (KDD 1999).

Theoretical Results

- A set of nodes P is a δ-false positive, if the Jaccard affinity in the original data is less than $\min_{i \in P} T(i)$, but it is reported as a valid local community by the min-hash approximation with an estimated affinity of at least $\min_{i \in P} T(i) \cdot (1 + \delta)$.

- A set of nodes P is a δ-false negative, if the Jaccard affinity in the original data is larger than $\min_{i \in P} T(i)$, but it is not reported as a valid local community by the min-hash approximation scheme, since the estimated affinity is less than $\min_{i \in P} T(i) \cdot (1 - \delta)$.

Theoretical Results

- The probability that a given set of nodes P is a δ-false positive for a min-hash sample of size k is given by at most $e^{-\delta^2 \cdot k \cdot \min_{i \in P} T(i)/4}$.

- The probability that a given set of nodes P is a δ-false negative for a min-hash sample of size k is given by at most $e^{-\delta^2 \cdot k \cdot \min_{i \in P} T(i)/2}$.

Consolidating Local Communities

- The local communities determined in the previous section need to be consolidated into a final set of compact communities.

- The min-hash technique of the previous section will create a large number of overlapping communities which need to be consolidated into a coherent set of communities.

- We use a two phase approach:
  - The first phase pieces together local communities in order to create the cores of the locally relevant communities.
  - The second phase then re-constructs these cores in a more comprehensive way with an iterative approach.

Experimental Results

- Tested on real and synthetic data sets
  - DBLP, Condensed Matter Physics (arxiv) and synthetic data set generated by RMAT
- Tested with effectiveness measures
- Tested with case studies for real data set

Effectiveness Measures

- We remove some of the nodes and their incident edges from the data, and perform the clustering on the remaining data set.

- We test how well their links relate to the different clusters which were created without the use of these nodes.

- The *dominant purity* p_i of node i is defined as the fraction of the links of node i which are incident on its dominant community.

- We define the *dominant interest ratio* I_i of a node i as the ratio of the dominant purity of node i to the fraction of the total number of network nodes which are contained in the dominantly linked community of node i.

Case Studies

- For the DBLP data set, the Newman algorithm created two very large communities each of which contained about 20% of the DBLP authors (400 communities in total).

- One of these large communities generated by the Newman algorithm contained the following set of authors:
  Jiawei Han, Mani Srivastava, Rajeev Alur, Donald Towsley, Barbara Liskov ...

- Mixes communities of researchers from different areas

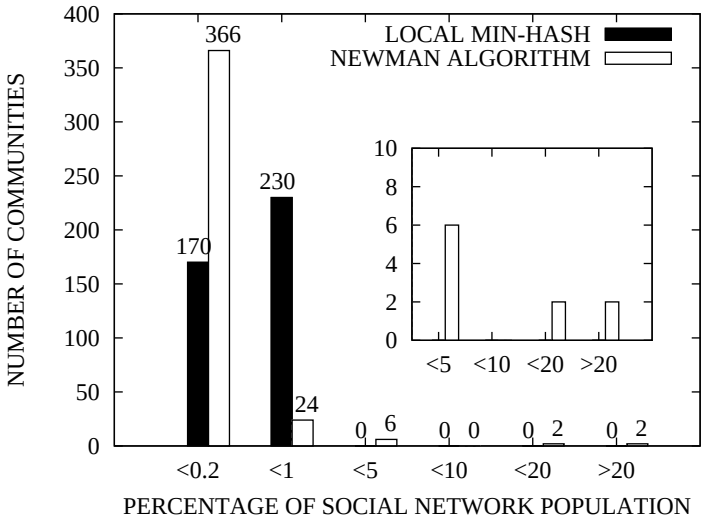
Case Studies

- Each of these authors was placed in a different community by the local community detection algorithm.

- The community for Jiawei Han contained less than 1% of the total authors, and contained the following individuals:
  Jiawei Han, HongJiang Zhang, Lei Zhang, ChengXiang Zhai, ...

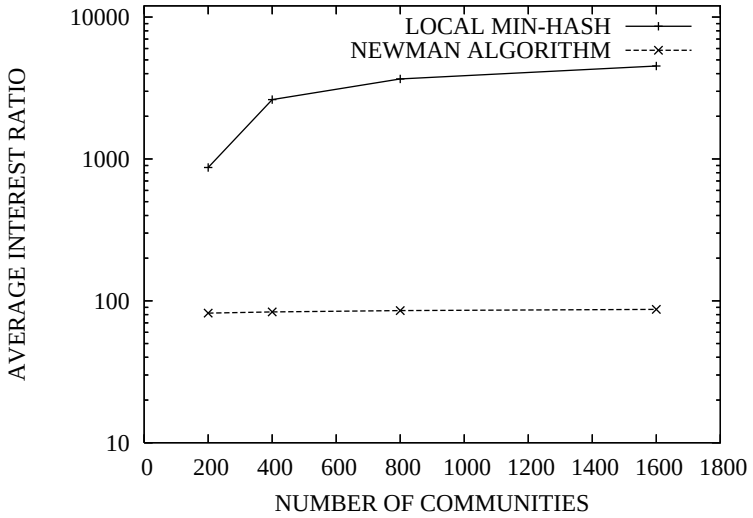
- Much more coherent set of researchers

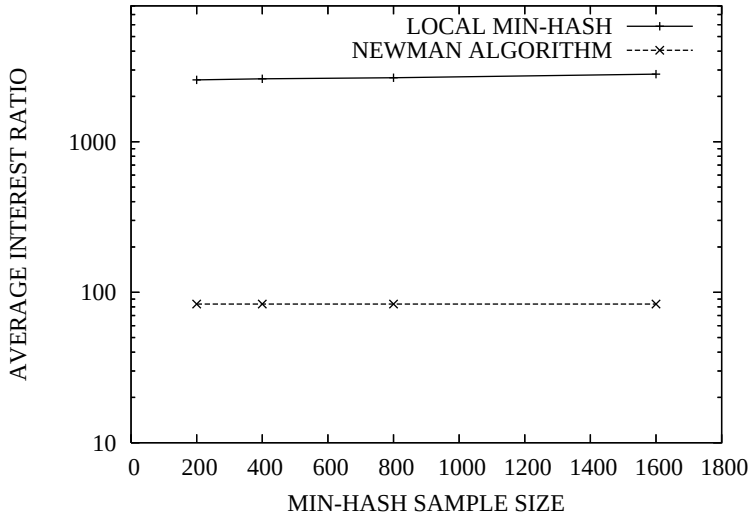
Distribution of points in clusters



- Distribution of data points in clusters

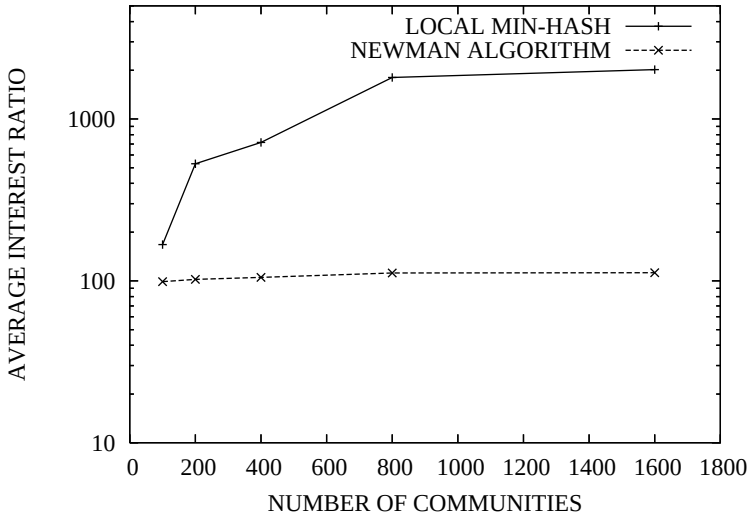
Effectiveness Results on DBLP Data Set

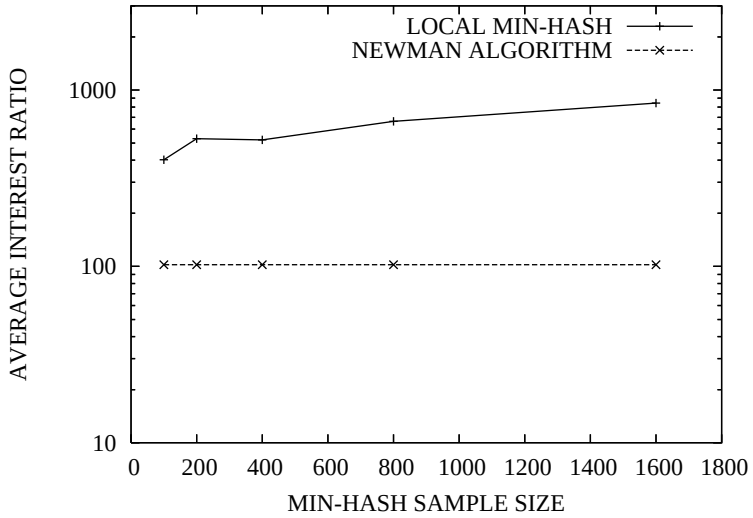




- Increasing min-hash sample size and number of communities

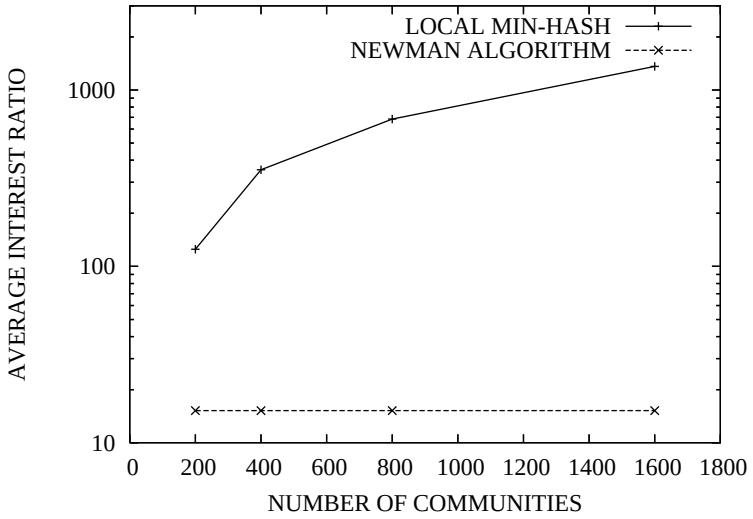
Effectiveness Results on Condensed Matter Data Set

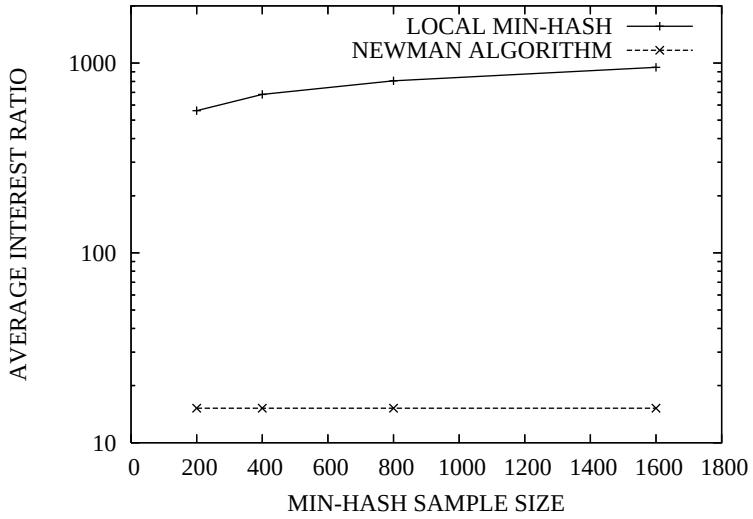




- Increasing min-hash sample size and number of communities

Effectiveness Results on Synthetic Data Set





- Increasing min-hash sample size and number of communities

Conclusions and Summary

- New method for local community detection in heterogeneous networks
- Uses a min-hash scheme for local community detection
- Determines more robust communities than a global approach