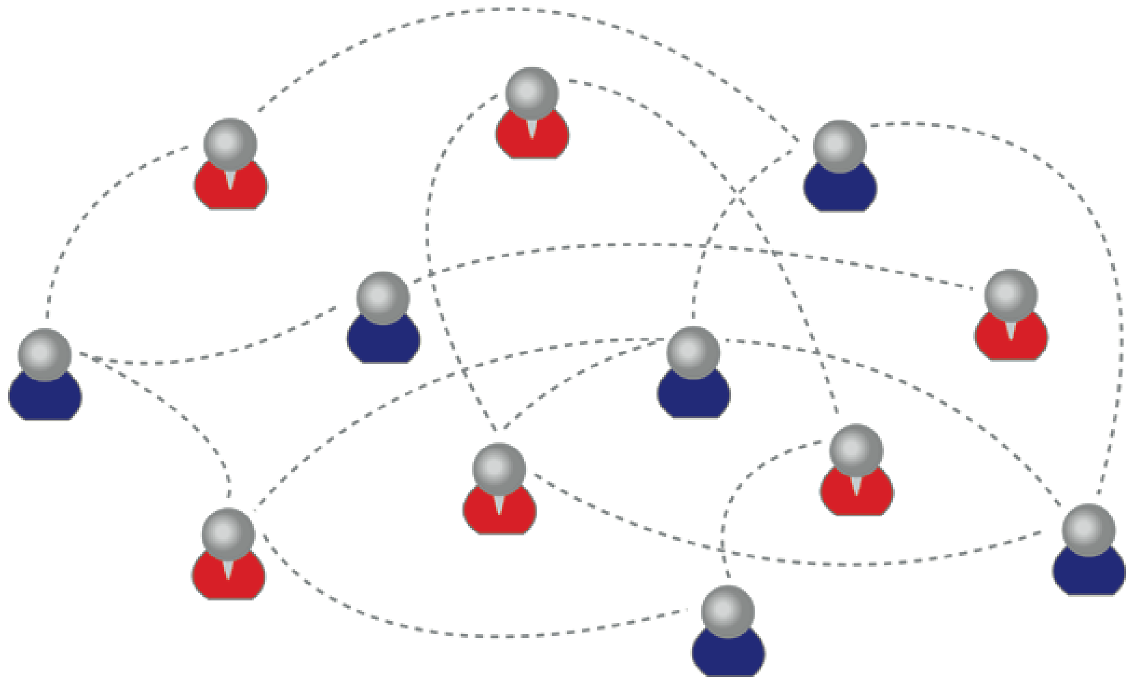

Finding Overlapping Communities in Social Networks: Toward a Rigorous Approach

Sanjeev Arora   Rong Ge   Sushant Sachdeva   Grant Schoenebeck

Presented by Eldad Rubinstein   July 4, 2012

Introduction

- What is a **community in a social network?**
  - a group of nodes more densely connected with each other than with the rest of the network
- Communities **overlap** each other
- Direct approach → **NP-hard** problems
- Heuristic or generative model approach → **egg & chicken** problem
- Instead: Assumptions are based on **ego-centric networks**
  - Studied in sociology
  - Suggested algorithms also have ego-centric analysis feel

Assumptions

0. Each person participates in **up to *d* communities**
   - *d* is constant or small
1. **Expected degree model**
   - Each node *u* in community *C* has an *affinity* $p_u \in [0,1]$
   - The edge *(u,v)* exists with probability $p_u p_v$
2. **Maximality** with gap ε
   - If for *u,v* $\in C$, *(u,v)* exists with probability α,
     then *w* $\notin C$ has edges to $\leq \alpha - \varepsilon$ fraction of nodes in *C*
3. **Communities explain** γ **fraction** of each person ties

First Step: Communities are Cliques

- Another Assumption: $\forall C:\ \delta k \leq |C| \leq k$
- Output each community with prob. $\geq 2/3$
  - in time $O(nk) \cdot 2^{\tilde{O}(\log^2 d)}$
- Algorithm Description
  1. Pick **starting nodes** uniformly at random
  2. For each starting node *v*, **randomly sample** $S \subseteq \Gamma(v)$
  3. Look at **cliques *U*** in *G(S)*
  4. Let *V'* be the set of nodes in $\Gamma(v)$ which are **connected to all** nodes in *U*
  5. Return **high degree vertices** from *G(V')*

Communities are Dense Subgraphs

- **Setup 1:** $p_u = \sqrt{\alpha}$
  - Find each community
    - With high probability over *G* randomness
    - With prob. 2/3 over algorithm randomness
    - In time $O(nk) \cdot 2^{\tilde{O}(\log^2 d)}$
- **Setup 2:** $p_u \geq \sqrt{\alpha}$
  - Need to **loop over** all $S \subseteq \Gamma(v)$ of size *T*
    - Sample $G(\Gamma(S))$ for each *S*
  - Worse running time: $O(n \cdot (kd)^T) \cdot 2^{\tilde{O}(\log^2 d)}$

Communities with Very Different Sizes

- Sampling may **miss small communities**
  - So previous ideas will not work
- **Definition:** *A* is a $(\alpha, \alpha - \varepsilon)$-*set* if
  - Nodes in *A* have edges to $\geq \alpha$ fraction of nodes in *A*
  - Outside nodes have edges to $\leq \alpha - \varepsilon$ fraction of nodes in *A*
- **Algorithm** (assuming $p_u \geq \sqrt{\alpha_{min}}$)

  1. For $\alpha = 1$ downto α_{min} step $-\varepsilon/4$

     1.1. For all sets of nodes *S* of size *T*

       1.1.1. *U* = {*v*: $\geq \alpha - \varepsilon/4$ fraction of its edges are to *S*}

       1.1.2. Return *U* if it is a $(\alpha, \alpha - \varepsilon/2)$ set
- **Running time:** $n^{C \log kd}$ (not polynomial)

Cliques with Very Different Sizes

- Looking for a **polynomial** algorithm for **cliques**
- **Extra assumptions** are needed:
  - Distinctness: For $u \in C$, at least a constant factor of *C* does not lie in any other community containing *u*
  - Duck assumption
  - Small communities are distinguishable from "noise" edges
- Polynomial algorithm description
  - Find large cliques first (sampled easily), then ignore their edges
  - Extra assumptions ensure smaller cliques can be found

Relaxing the Assumptions

- **Expected degree model** assumption can be relaxed if:
  - The following are **concentrated** near their expectation:
    - **# of edges** from any node *u* to any community *C*
    - **Degree** of each node
    - **Intersection** of two nodes in a community
- **Gap assumption**
  - Can be relaxed if:
    - $\forall C : \delta k \leq |C| \leq k$
    - Communities are cliques or $p_u = \sqrt{\alpha}$
  - The returned communities will be close to the real ones

Sparser Communities

- Different assumptions
  - *(u,v)* exists with probability $B/\sqrt{k}$ (where $|C| = k$)
  - All edges belong to some community
  - Communities intersection size is limited
- Transform *G* to a dense graph *G'*
  - Nodes are the same
  - *(u,v)* exists in *G'* iff they have $\geq B^2/2$ length-2 path in *G*

Summary

case no.	extra / different assumptions?	probability of edges in communities	communities sizes must be similar?	running time
1	No	Cliques	Yes	Polynomial
2	No	$p_u = \sqrt{\alpha}$	Yes	Polynomial
3	No	$p_u \geq \sqrt{\alpha}$	Yes	Polynomial
4	No	$p_u \geq \sqrt{\alpha}$	No	Quasi-Poly
5	Extra	Cliques	No	Polynomial
6	Different	Sparse	Yes	Polynomial

Areas of Possible Further Research

- Releasing the assumptions in more cases
  - Expected degree model assumption
  - Maximality (gap) assumption
- Polynomial algorithm for dense communities with different sizes
- Fast implementation using heuristics
- Testing on real-world data
- Adapting the algorithms to a dynamic setting


digg
YouTube
Broadcast Yourself
upcoming.org
last.fm
the social music revolution
flickr
BETA
myspace.com
Technorati
twitter
facebook
MyBlogLog
slideshare
jaiku

Questions?