CS3230R

Community Detection in graphs

A review paper by

Santo Fortunato

Davin Choo

Outline

- 15. Testing (Pg 145 ~ 154 in review paper)
  - **15.1 Benchmarks**
  - 15.2 Comparing partitions: Measures
  - ~~15.3 Comparing algorithms~~

Just some analysis and comparisons of algorithms discussed in paper

Overview Remark

- No standardised testing method agreed upon
  - Why? Because no standardised definition of "community" to begin with
- Usually use popular methods
- Consequence:
  No standardised benchmark → Any algorithm is good "in some sense"/benchmark of their own

Popular benchmarks

- Planted L-partition model
(supposed to be small lettered l, but L for readability)
- LFR benchmark
(more realistic model of real world graphs)

Planted L-partition model

- Computer-generated
  - Partition graph into L groups with g vertices each
  - Intra-group, add edge with probability $\mathbf{p_{in}}$
  - Inter-group, add edge with probability $\mathbf{p_{out}}$
- Properties
  - Each group (as a subgraph) is a Erdos-Renyi graph with connection probability $\mathbf{p_{in}}$
  - Average degree of a vertex $\langle k \rangle = p_{in}(g-1) + p_{out}g(L-1)$

Planted L-partition model

- Common to use (Gained status of “standard benchmarks”):
  - Girvan and Newman set **L = 4, g = 32, <k> = 16**
    - Note: $\mathbf{p_{in}}$ and $\mathbf{p_{out}}$ are hence dependent on each other
  - Let $\mathbf{\langle k \rangle = z_{in} + z_{out}}$
    [indicates expected internal/external degree of a vertex]
  - $\mathbf{z_{in} = p_{in}(g\text{-}1) = 31p_{in}}$
  - $\mathbf{z_{out} = p_{out}g(L\text{-}1) = 96p_{out}}$
  - Able to detected the planted partition up until $z_{out} \approx 12$
    (i.e. $z_{in} \approx 16 - 12 = 4$, $p_{in} = p_{out} = 1/8$; We get a truly random graph)

Planted L-partition model

- Usage
  - Build a few graphs for a fixed z_{out}
  - Compute average similarity (refer to 15.2) between solution of method and built-in/planted solution
  - Iterate on different values of z_{out}
  - Plot graph (X-axis = z_{out}, Y-axis = similarity)
- Usually, perform well on low z_{out} and start to fail when z_{out} approaches 8

Modifications to Planted L-partition model

- Fan et al. [Keep p_{in} and p_{out} independent]
- Brandes et al. [Gaussian random partition generator]
- Lancichinetti et al. [LFR benchmark]

LFR benchmark

- Assume distributions of degree and community size are power laws, with exponents τ_1 and τ_2, respectively
- $\forall$v, v shares (1 - μ) edges with v′ in same community
  - Mixing parameter μ : $0 \leq \mu \leq 1$

LFR benchmark

- Building steps

1. Pick a sequence of community size using τ_2
2. $\forall v_i$, generate k_i (degree of v_i) using τ_1
   Set internal degree of v_i to $(1 - \mu)k_i$
   Set external degree of v_i to μk_i
3. Randomly connect vertices within communities until all internal edges are filled up
4. Randomly connect vertices across communities until all external edges are filled up

LFR benchmark

- Numerical tests show that building is O(m), where m = #edges in graph
- A. Lancichinetti, S. Fortunato
  [Extend LFR benchmark to directed and weighted graphs with overlapping communities]
- Free download link for software to create LFR benchmark graphs:
  http://santo.fortunato.googlepages.com/inthepress2

Other benchmarks
(inspired by Planted L-partition model)

- Bagrow [Graphs with power law degree distribution]
- D.J. Watts [Relaxed Caveman graphs]
  - Originally used to explain clustering properties of social networks
- Arenas et al. [Embedded hierarchical structure]
- Guimera et al. [Bipartite graphs]
- Sawardecker et al.
  [General model, accounts for possibility of cluster overlap]

Outline

- 15. Testing (Pg 145 ~ 154 in review paper)
  - 15.1 Benchmarks
  - **15.2 Comparing partitions: Measures**
  - ~~15.3 Comparing algorithms~~

Popular measures

- Girvan and Newman
  - **Fraction of correctly classified vertices**
- Others can be divided into 1 of 3 categories:
  - Pair counting
  - Cluster matching
  - Information theory

Popular measures

- Girvan and Newman
  - **Fraction of correctly classified vertices**
- Others can be divided into 1 of 3 categories:
  - Pair counting
  - Cluster matching
  - Information theory

Some measures

(__) is equation number in the paper

- Let $\mathcal{X}$ and $\mathcal{Y}$ be 2 partitions of graph G
  - Wallace's 2 indices $\mathbf{W_I}$ and $\mathbf{W_{II}}$ (87)
  - Rand index $\mathbf{R}(\mathcal{X}, \mathcal{Y})$ (88)
  - Mirkin metric $\mathbf{M}(\mathcal{X}, \mathcal{Y})$ (89)
  - Jaccard index $\mathbf{J}(\mathcal{X}, \mathcal{Y})$ (90)
  - Classification error $\mathbf{H}(\mathcal{X}, \mathcal{Y})$ (91)
  - normalized Van Dongen metric $\mathbf{D}(\mathcal{X}, \mathcal{Y})$ (92)
  - normalized mutual information $\mathbf{I_{norm}}(\mathcal{X}, \mathcal{Y})$ (93)
  - Variation of information $\mathbf{V}(\mathcal{X}, \mathcal{Y})$ (94)
  - meet $\mathscr{M}$ (95)
  - Relative overlap $\mathbf{s_{ij}}$ (97)

Questions?

- Slides will be made available for reference