Structural Properties and Generative Model of Non-giant Connected Components in Social Networks

Jianwei Niu

Beihang University

niujianwei@buaa.edu.cn

Beijing, Aug. 13 2015

SCIENCE CHINA PRESS

北京航空航天大學
BEIHANG UNIVERSITY

Outline

- ➢Motivation
- ➢Related work
- ➢Datasets
- ➢Definitions
- ➢Observations
- ➢Proposed model
- ➢Summary

Motivation

- Real-world networks often contain a Giant Connected Component (GCC), but also many non-giant connected components, or Disconnected Components (DCs).
- Most previous research were conducted on the giant component implicitly or explicitly. Properties of non-giant components are seldom mentioned.
- Understanding the evolution of non-giant connected components may help us in:
  - Graph generation and simulation
  - Predicting future state of networks

The problem

- What can we say about the evolution of non-giant connected components?
  - □ What information can we get by studying the local structure of the disconnected components in a network?
  - □ Can we design a model to reproduce the observed properties?

Related Work

- Network patterns
  - Densification power law (Leskovec. et al. KDD 05)
  - Shrinking diameter (Leskovec. et al. KDD 05)
  - Gelling point (McGlohon et al. KDD 08)
  - Middle regions, star-shaped (Kumar et al. KDD 06)
- Generating models
  - Erdos-Renyi, Barabasi-Albert, Preferential Attachment
  - Community-guided attachment and Forest Fire models (Leskovec et al. KDD 05)
  - Butterfly model (McGlohon et al. KDD 08)

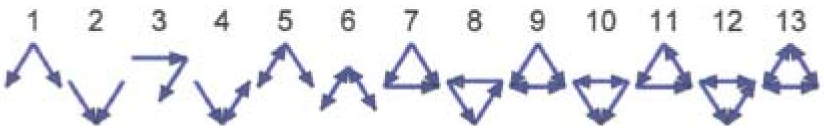
Datasets

- For the all three datasets that we use in our study, the facebook and twitter datasets are publicly available.
- We obtain the renren dataset through our collaborations on research with the Renren Corporation.

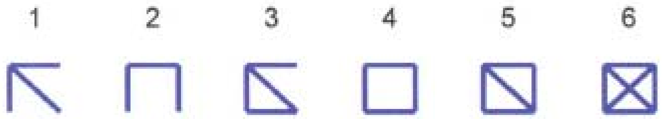
Name	|N|	|E|	Type
renren	516, 765	6, 866, 141	Undirected
facebook	4, 039	88, 234	Undirected
twitter	81, 306	1, 768, 149	Directed

Definitions

- Significance Profile (SP):
  - To calculate the SP of a network, the network is compared to an ensemble of randomized networks with the same degree sequence
  - The significance profile means the importance of 13 different triads for a directed graph (Fig. a) and 6 different quadruples for an undirected graph (Fig. b)



(a)



(b)

Definitions

- Significance Profile (SP):
  - To calculate the SP of a graph, we first randomize the graph, and for each of the 13 triads for a directed graph and the 6 quadruples for an undirected graph, we compute the Z_i as follows,

$$Z_i = \frac{Nreal_i - < Nrand_i >}{std < Nrand_i >}$$

where the $Nreal_i$ means the times some triad or quadruple appears in a graph, and the $Nrand_i$ means the times some triad or quadruple appears in the randomized graph. $< Nrand_i >$ and $std < Nrand_i >$ denote the average value and standard deviation of $Nrand_i$ respectively.

  - Then we normalize the array Z to length 1, and get the SP_i:

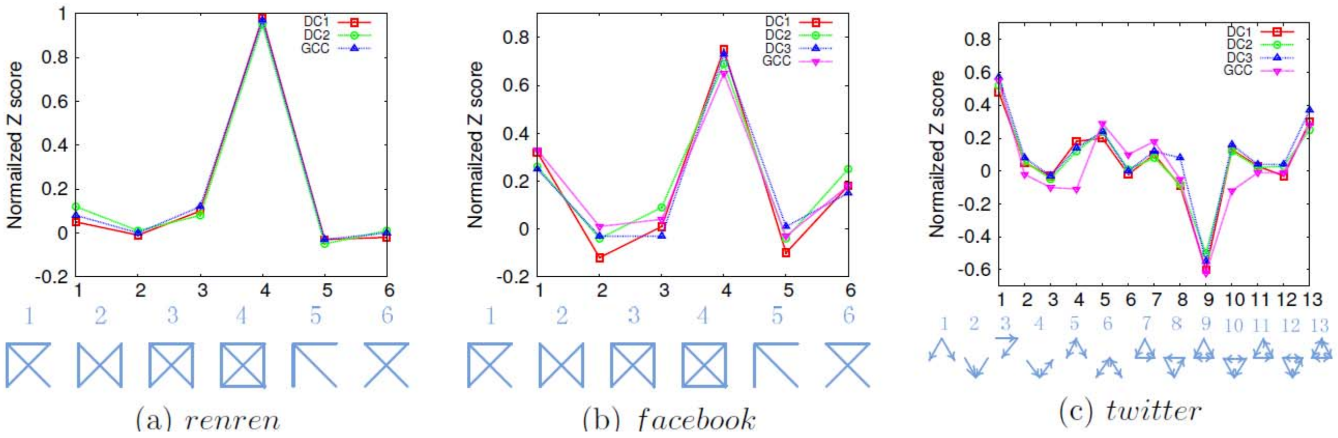
$$SP_i = \frac{Z_i}{(\sum Z_i^2)^{1/2}}$$

Observations

- The significance profile (SP) of GCC and DCs
  - The adding process of nodes and edges usually follows a same or similar pattern in the same network.
  - So we want to find if there exists any similar local structure between the GCC and DCs.
  - To study the SP of the GCC and DCs, we first divide the different components of a graph into the GCC and DCs as subgraphs. Then we calculate the significance profile (SP) of these components.

Observations

- The significance profile (SP) of GCC and DCs
  - We can observe that in these datasets, the GCC and the DCs share the similar SP_i. So we can draw the conclusion that in these networks there are some local structural similarities between the GCC and DCs.



The significance profile of subgraph for different components of a graph.

Observations

- The significance profile (SP) of GCC and DCs
  - In our study, we also find that in some graphs the triad significance profile (TSP) or quadruple significance profile (QSP) of the DCs are quite different from the GCC.
  - These graphs are often the topology of “real” networks like collaboration networks. The underlying reason may be that these disconnected components are set for “especial” purpose, so they are separated from others.
  - But in networks like topology of relationship on SNS, citation network, website links, we can always see the similar TSP or QSP between the GCC and DCs. We may explain it as the nodes in these networks are “un-organized” individuals who just make their decisions randomly, so there are not so many “special” DCs in the network.

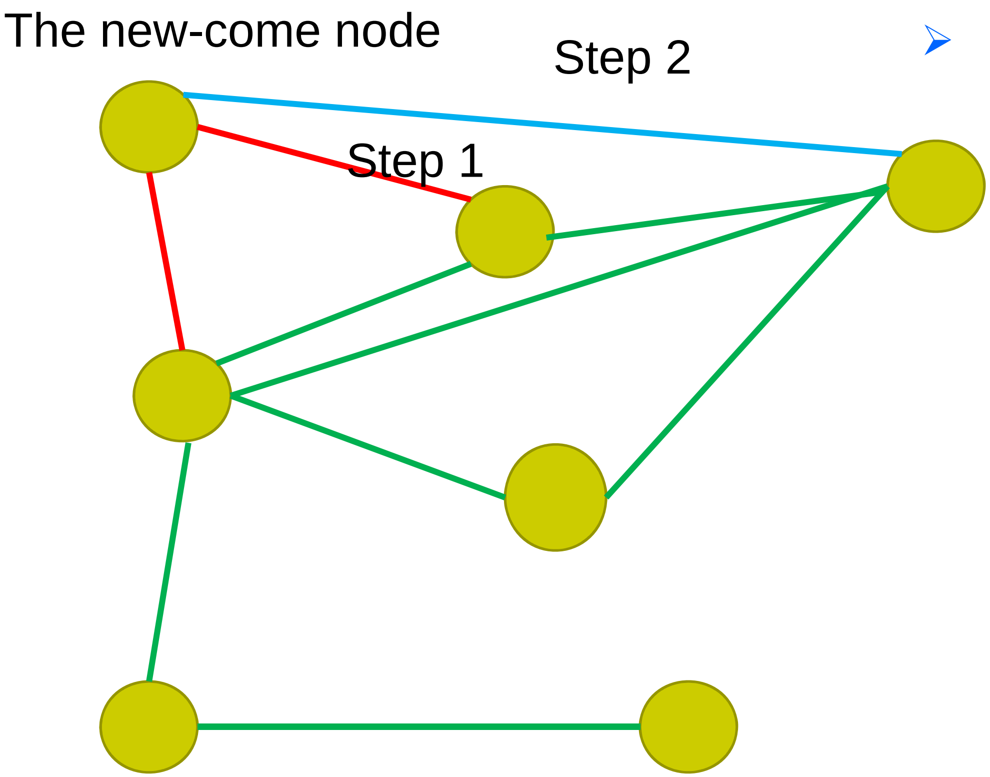
Proposed Model

- Intuitive ideas:
  - Two persons who don’t know each other are likely to become friends if they share a lot of similarities.
  - Two persons who don’t know each other are likely to become friends if they share a lot of friends.
- Desired properties:
  - The decaying curve in the longevity and size distribution of non-giant connected components.
  - Local structural similarities between the GCC and DCs.

Proposed Model

- A new-come node N_i is born with a n-dimensional social property vector p_i=(v_1,v_2,…,v_n)
- The social-distance between N_i and N_j is defined as $D_{i,j}$=||p_i-p_j||
- The more similarities that nodes i and j share, the smaller value of $D_{i,j}$ is
- The bigger value of $D_{i,j}$ is, the less possible the edge between nodes i and j will be created with.

Proposed Model



- We start with a network G = (V, E), where |V| = 1 and E = ∅, and the newcomer nodes arrive one by one. The joining of each node includes two steps.

(1) Edge generation based on nodes' similarities. Wher a node N_i arrives, it will create connections with some existing nodes. The more similarities each node pair share, the higher possible the edge will be created.

(2) Edge generation based on mutual friends. Let S_A be the set of all the chosen nodes in step (1), and for each $N_j \in S_A$, we traverse the set of $Neighbor(N_j)$, i.e., the set of nodes that connect directly to N_j. For each $N_k \in Neighbor(N_j)$, the possibility of the creation of the edge $e(N_i, N_k)$, $P\{e(N_i, N_k)\}$, is computed as follows.

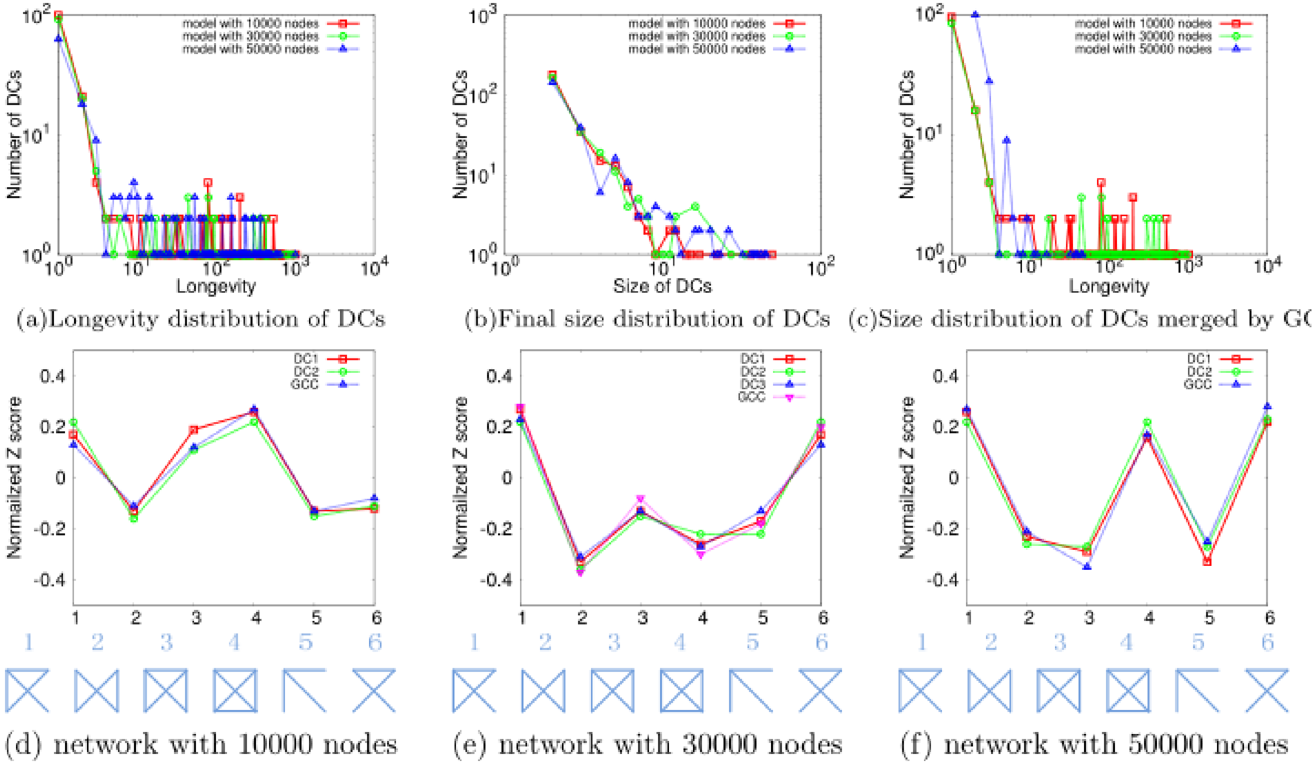
$$P\{e(N_i, N_k)\} \propto \frac{Neighbor(N_i) \cap Neighbor(N_k)}{Neighbor(N_i) \cup Neighbor(N_k)}$$

Proposed Model

- In this generative model, when a newcomer node forms edges to some existing nodes in the network, it can be connected to the GCC or existing DCs.
- A DC is claimed to be dead if it is connected to the GCC or merges with other DCs.
- In the former situation, the dead DC and the GCC become a new GCC.
- In the latter case, those DCs which merge with each other are merged into a new born DC.

Proposed Model

- Generate three networks with 10,000, 30,000 and 50,000 nodes respectively.



(a)Longevity distribution of DCs (b)Final size distribution of DCs (c)Size distribution of DCs merged by G(

(d) network with 10000 nodes (e) network with 30000 nodes (f) network with 50000 nodes

Results for each run of our model

Summary

- Observations on:
    - Structural Properties of Components
        - ✓ There exists a similarity in the local structure between the giant connected component and disconnected components in social networks
- A generating model
    - Decaying trend of components longevity and size distribution curve
    - Local structural similarities between the GCC and DCs

Thanks very much for your attention!

Any Questions?