A Randomized Sampling Algorithm based on Triangle for Community Extraction in Graphs

Yanting Li*, Ying Huo

School of Information Science and Engineering

Shao Guan University

yanting8015@sgu.edu.cn

Introduction of Presenter

Yanting Li

- 2009--2011 Kyushu Sangyo University (Japan) Master degree in computer science
- 2011--2015 Kyushu Institute of Technology (Japan) Ph.D in computer science
- 2015--2017 Shao Guan University (P.R.China) Assistant professor
- 2017--2018 Freie University (Germany) Visiting Scholar
- 2018--Present Shao Guan University (P.R.China) Associate professor

June, 2013 in Lisbon

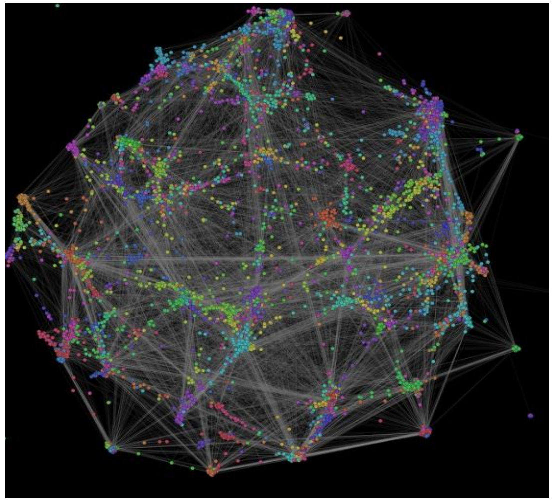
Introduction of Research Projects

- Currently, my research group is working on three projects
  - Query-focused keywords extraction by adopting tree search for document automatical abstraction. This project is funded by the Science and Technology Administration of Shaoguan.
  - An unsupervised learning approach of graph-based X-ray photograph prosessing for tumor diffusion characteristics analysis. The key idea of this approach is the computation of nodes shifting in the set of X-ray photographs so that the spreading of tumor can be predicted. This project is funded by the Natural Science Foundation of Guangdong Province.
  - An approach of compressed query algorithm based on LFB Storage structure for documents classification by extracting key-sentences.

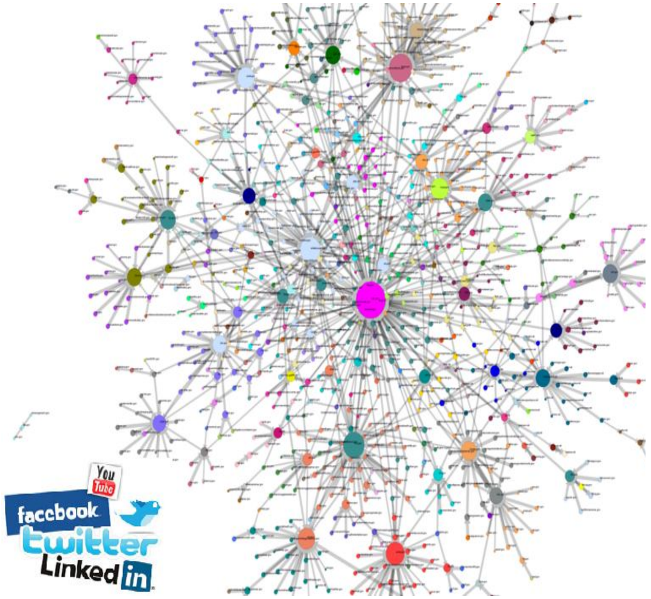
Outline

- **Background**
- **Key idea of the randomized sampling algorithm**
- **Mainframe of the randomized sampling algorithm**
  - Node coloring (generation of random values)
  - Edge sampling
  - Community Extraction
- **Experimental results**
- **Conclusion**

Background

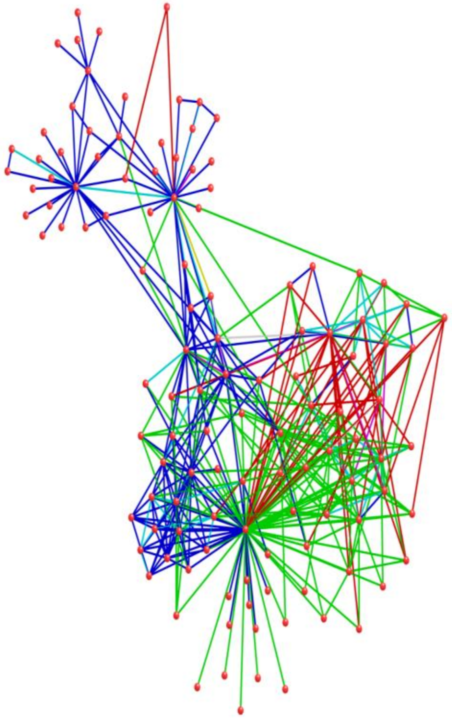


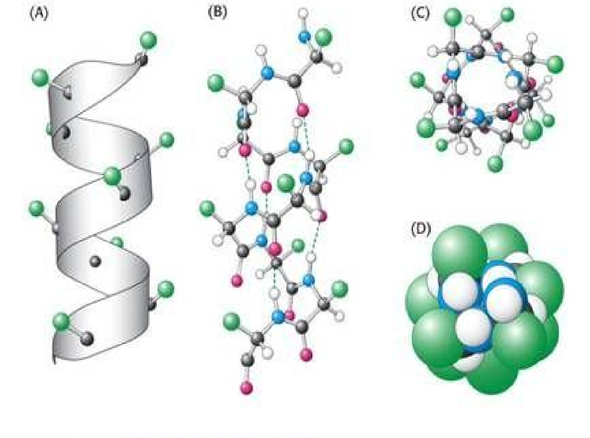
Webpage graph



Social network

Traffic network

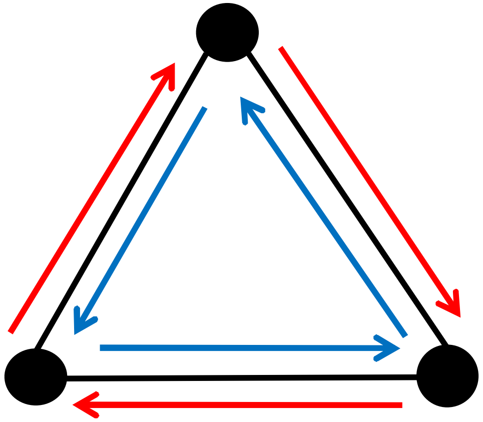


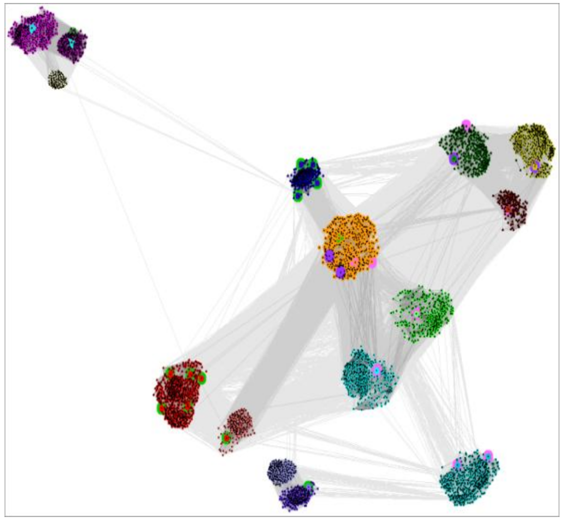


Protein structure

Background

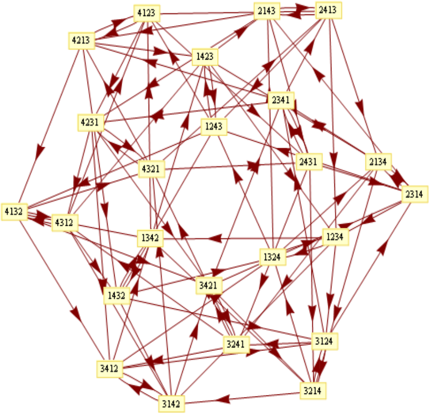
Clustering coefficient and transitivity ratio are two measurements frequently used in network analysis





Clustering

Transitivity



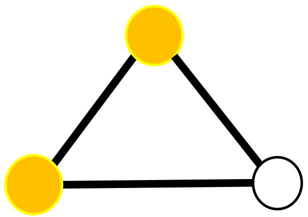
The increasing huge size of graph data with a complicated structure courses the high cost, mainly in time cost and memory cost

Contributions

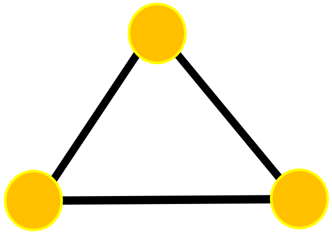
- The proposed algorithm is innovated by the following contributions
  - Each edge in graph *G* is selected based on the uniformly coloring of nodes with probability. Colors are denoted as real integer numbers, and randomly given to nodes
  - The third edge will be sampled if other two edges of a triangle are sampled. A triangle with three monochromatic edges is the smallest sampling unit. An extracted community must contain at least one triangle

Key idea of the randomized sampling algorithm

- The randomized sampling algorithm considers triangle as the samllest unit of community
- The key idea consist of two components
  - An edge is monochromatic if both its connected nodes in the same color
  - Correlate the sampling of edges that the third edge will be sampled if two edges of a triangle are sampled



An edge of a triangle is monochromatic that the triangle does not satisfy the parameter



Three edges of a triangle are monochromatic so that the triangle is sampled

Generation of random value

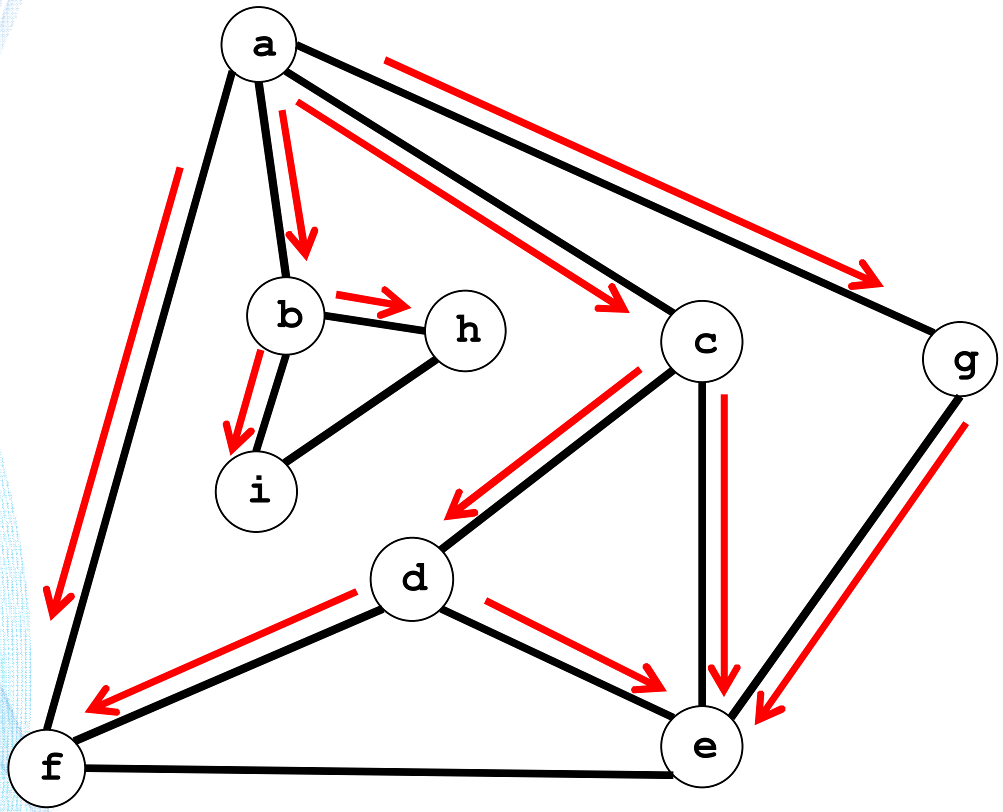
- Assume that the range of random values has finite expectations and variances mathematically. The generation of R_v can be gained.

$$x_n + 1 = (\frac{x_n^{\ 2}}{10^s})(\mathrm{mod} 10^{2s})$$

- The (X_n + 1) is an iterative operator, and (R_v + 1) is the random value R_v that needs to be generated every time. The s is the shifting of X_n square metre for generating new random value.

$$R_v + 1 = \frac{x_n + 1}{10^{2s}}$$

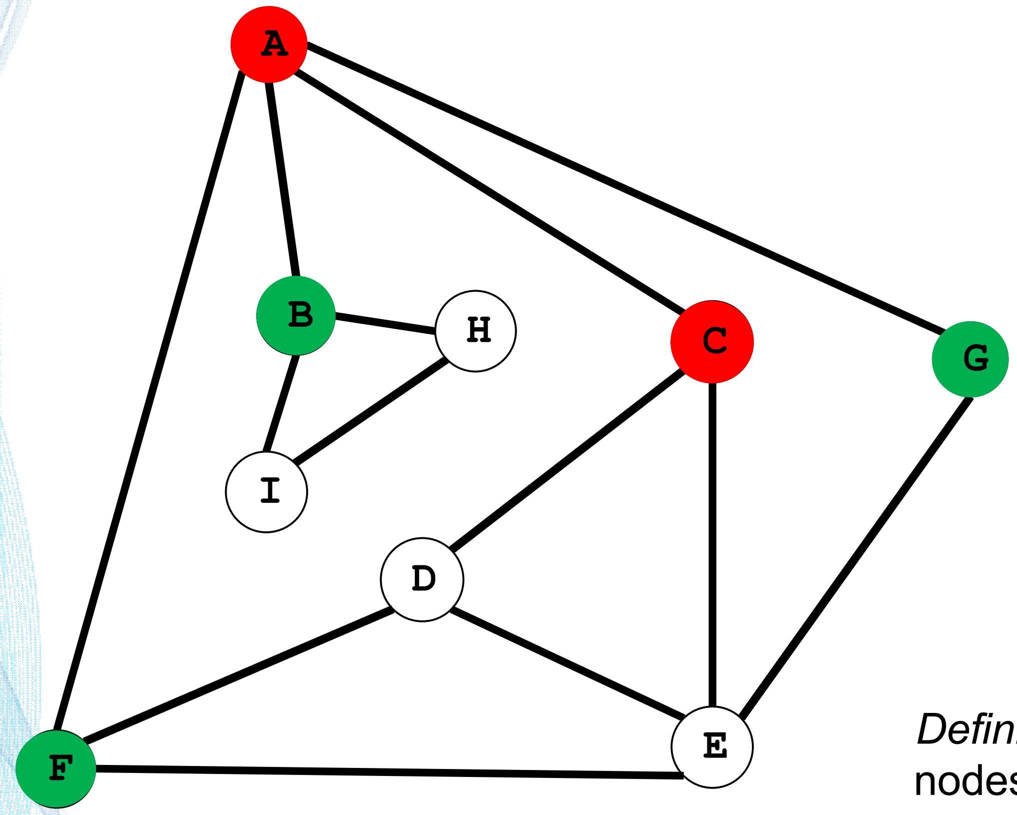
Mainframe of the randomized sampling algorithm



Adjacency Lists

a: b c f g
b: a h i
c: a d e
d: c e f
e: c d g
f: a d e
g: a e
h: b i
i: b h

Mainframe of the randomized sampling algorithm

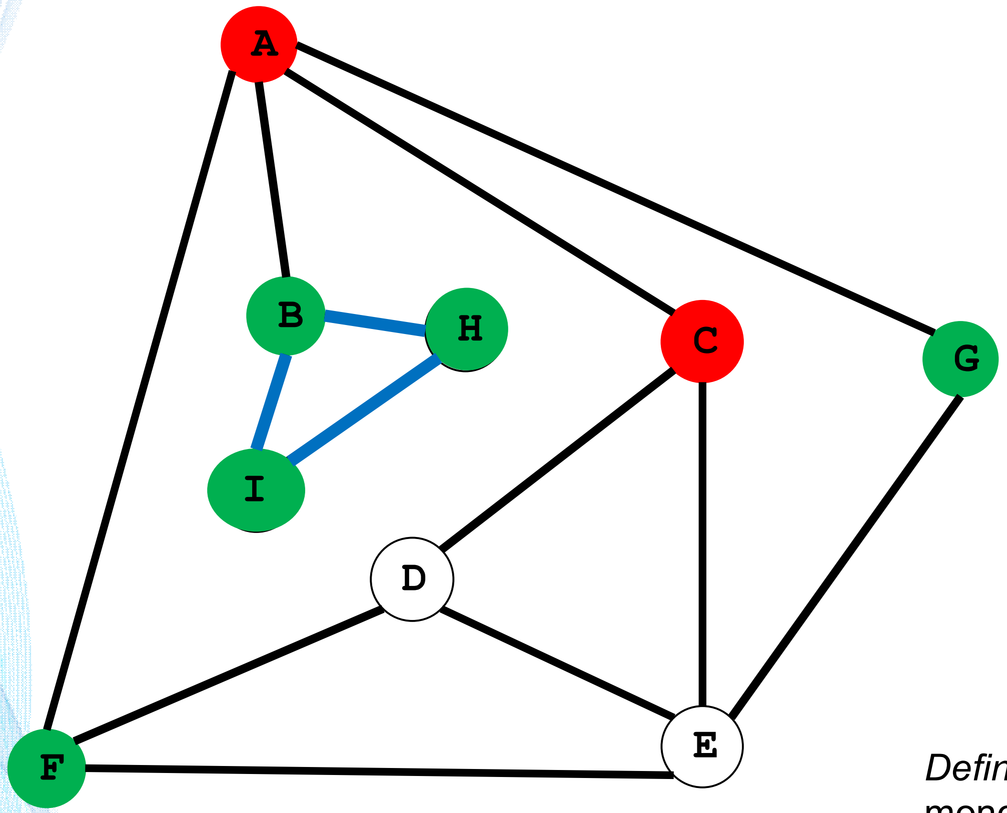


Function called array:

A: A-B A-C A-F A-G

Definition: The coloring of a nodes is defined as $cr(v, G)$ that R_v is uniformly given to each node where $0 < R_v < |n|$

Mainframe of the randomized sampling algorithm



Function called array:

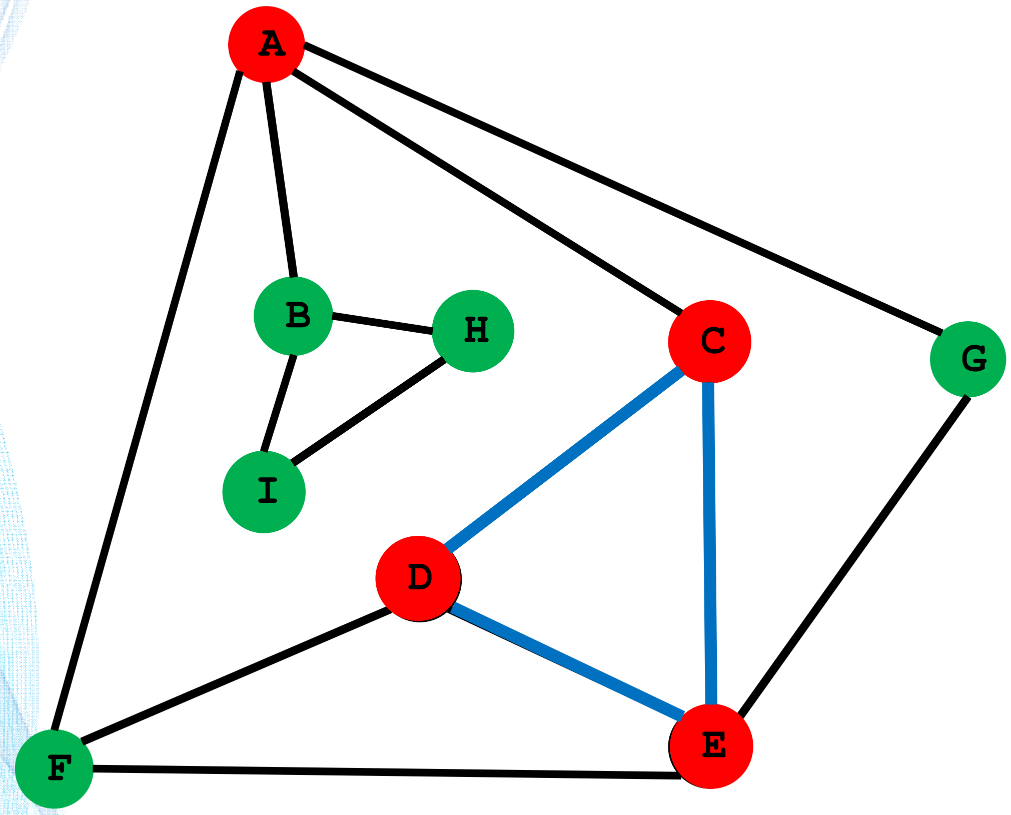
A: A-B A-C A-F A-G

B: B-A B-H B-I

Triangle： T_{BHI}

Definition: an edge is monochromatic if its two endpoints receive the same color where $R_i = R_j$.

Mainframe of the randomized sampling algorithm



Function called array:

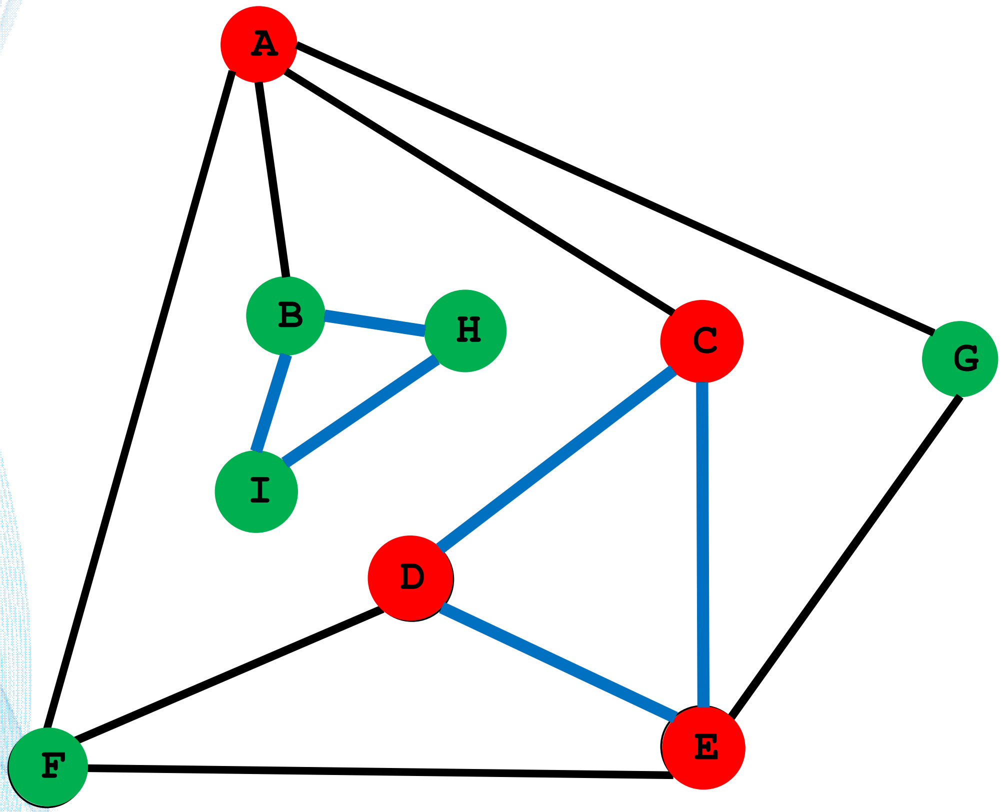
A: A-B A-C A-F A-G

B: B-A B-H B-I

C: C-A C-D C-E

Triangle：T_{CDE}

Mainframe of the randomized sampling algorithm



Function called array:

A: A-B A-C A-F A-G

B: B-A B-H B-I

C: C-A C-D C-E

D: D-C D-E D-F

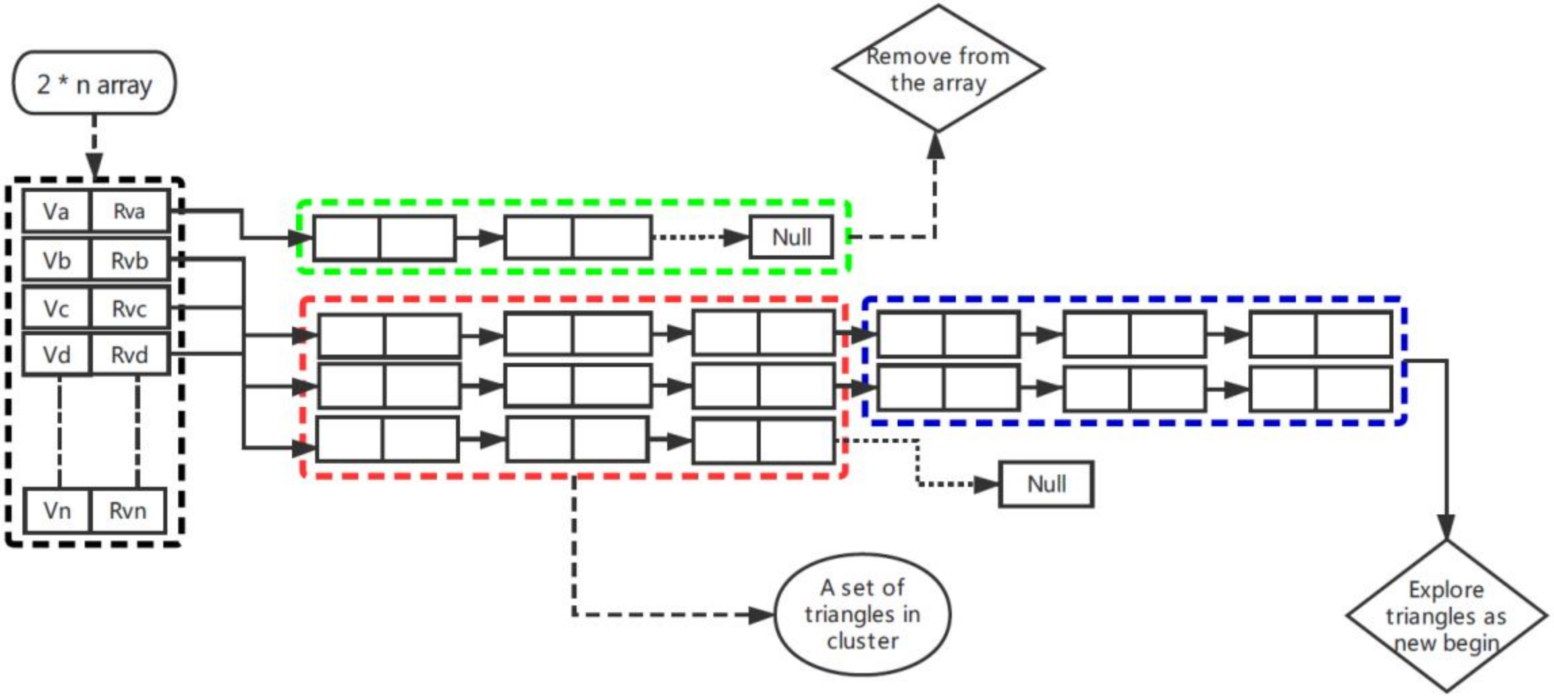
Triangle： T_{BHI} and T_{DEF}

Probability Analysis

- Two ways of sampling are considered as below
  - Global Sampling considers the probability of a triangle to be extracted from *G*. A tiangle that consists of three monochromatic edges $\{e_{ij}, e_{jk}, e_{ik}\} \in MONO_e$ is extracted as a smallest community of *G*
  - Local Sampling considers the probability of an edge *e* to be sampled as a monochromatic edge. For every two nodes *j* and *k* that connected by an edge $e_{jk} \in E_G$ receive the same color. Such that, the edge e_{jk} is monochromatic. The $Pr_{(jk)} = Pr_{(j)} \times Pr_{(k)}$

Implementation

- A (2 n) array list is employed for building the storage of all nodes in V_G and their corresponding random values



Features of datasets

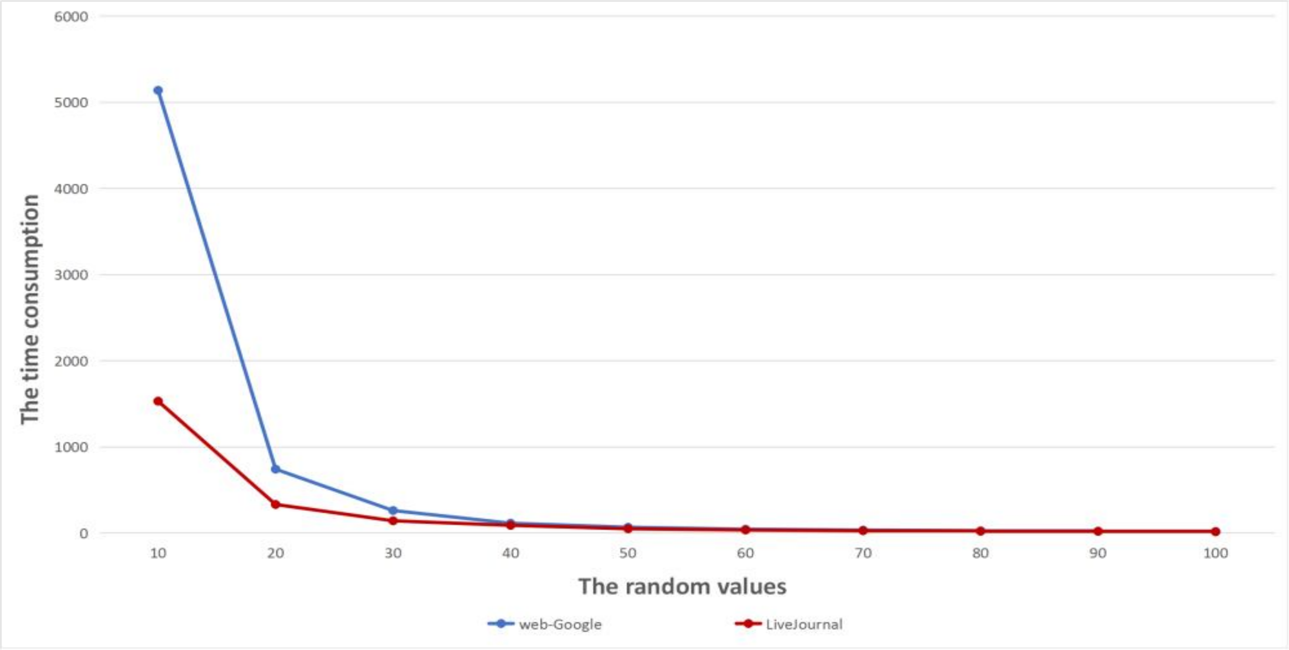
- Features of datasets for experiment

Name	Nodes	Edges	Description
Web-google	875,713	5,105,039	Road network of California
com-LiveJournal	3,997,962	34,681,189	Live-Journal online social network

- Environment
  - The program was ran on a machine with an Intel i7 2.3GHz CPU and 16GB RAM
  - Use g++ 4.1.2 compiler in Mac OS

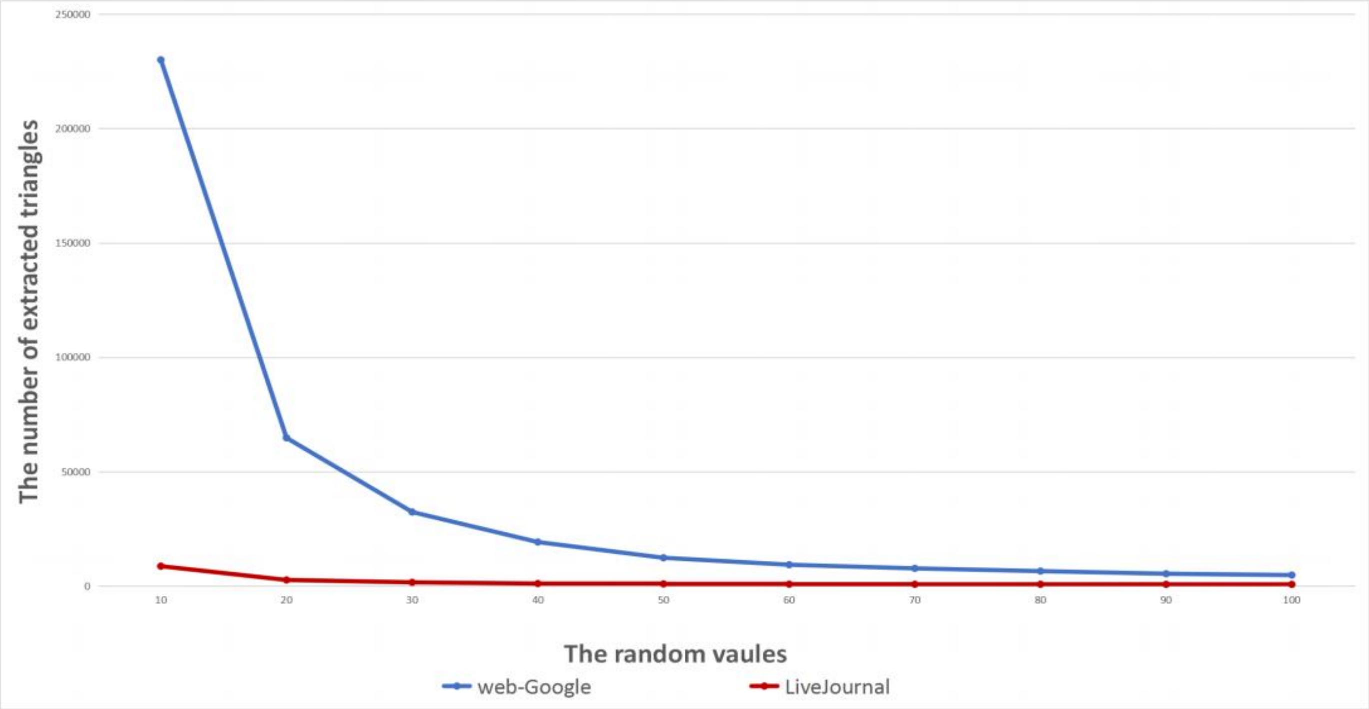
Experimental results

- This experiment records the time consumption in extracting communities
- With the increasing random value, the time cost for extracting communities decreases for less numbers of triangles are extracted



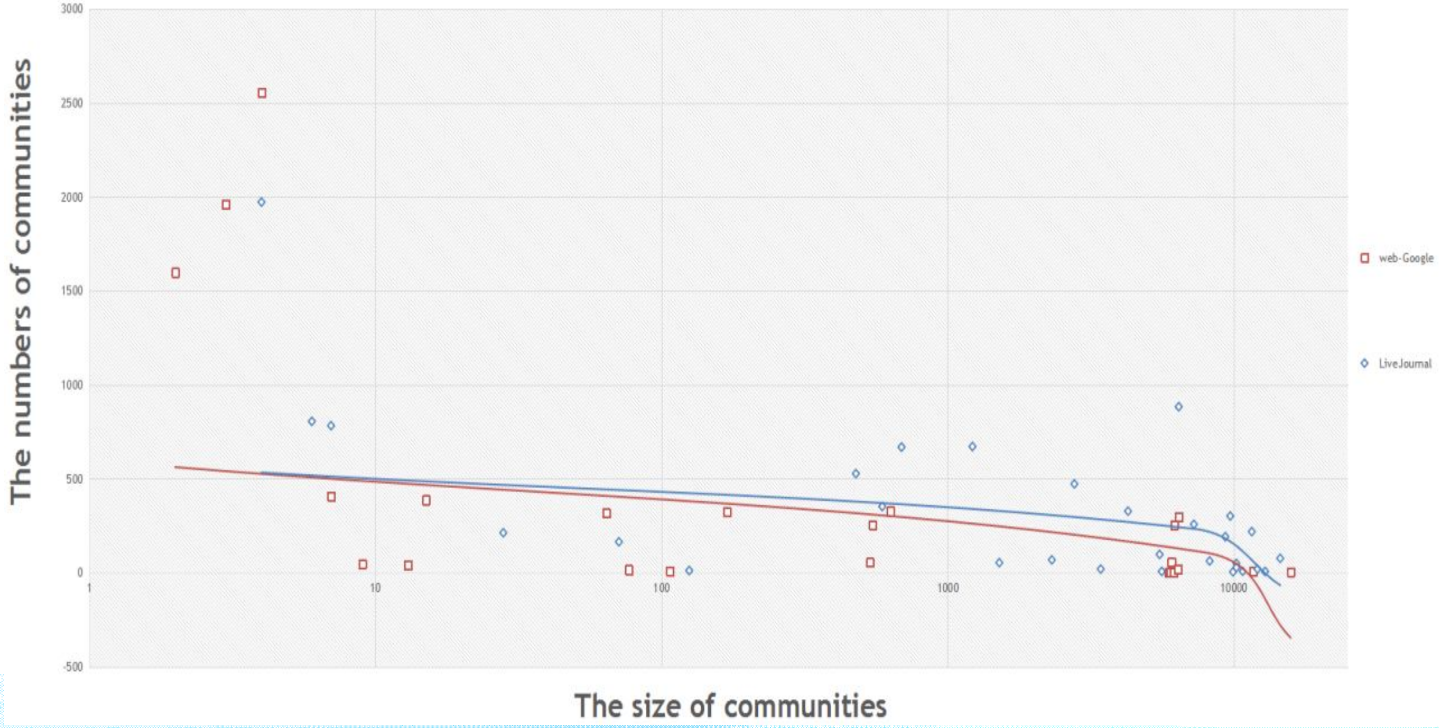
Experimental results

- This experimental result shows the relationship between the number of extracted triangles and the random values
- The numbers of extracted triangles decrease with the increasing of random value for the probability of an edge being sampled reduces



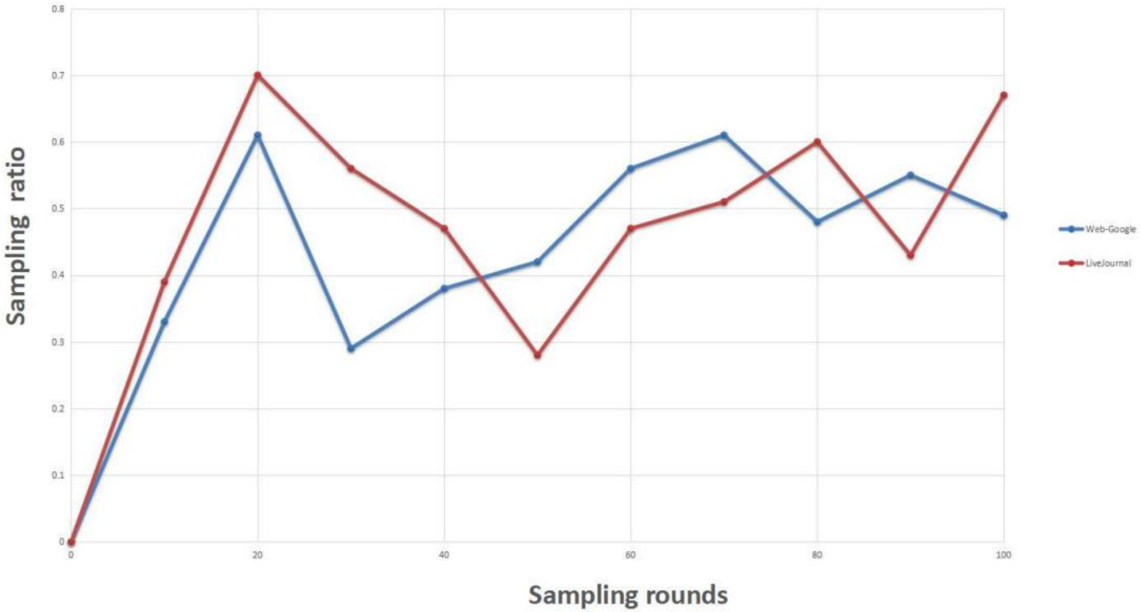
Experimental results

- This experiment proves the distribution of communities
- Small value of R_v leads to higher probability of an edge to be monochromatic. A few number of communities in large size are obtained
- Large value of R_v leads to smaller probability of an edge to be monochromatic. A large numbers of communities in small size are extracted



Experimental results

- This experimental results record the statistics of sampling ratio
- The results of each sampling can be regarded as a random variable $Var_{(s)}$
- Due to the unknown total number of communities and the unpredictabe number of extracted samples X in the set of communities s, the n rounds sampling results $\{X_1, X_2, X_3 \ldots X_n\} \in s$ can be considered as a set of random variable $Var_{(s)}$

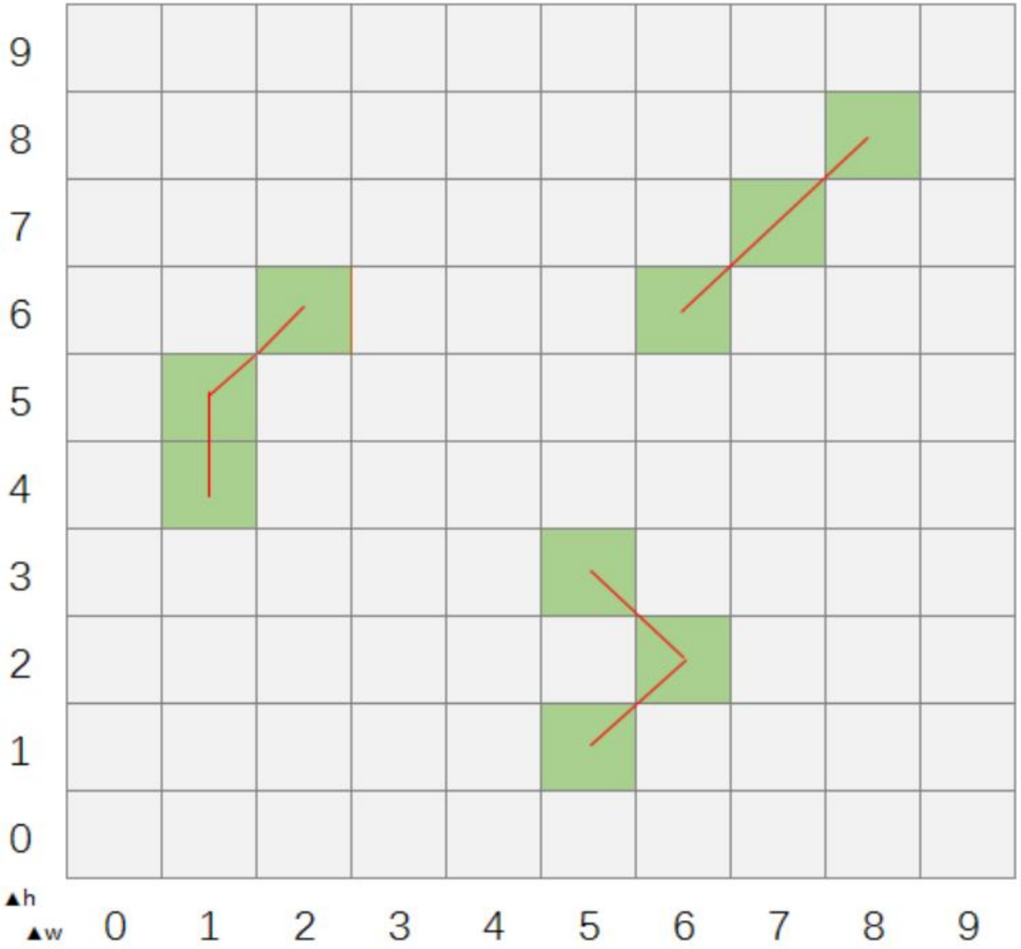


Experimental results

- A method of 2-Dimensional Grid for locating communities in *G* is proposed

- The red graph illustrates the path of searching communities

- The green blcoks indicate the location of triangles in *G*



Experimental results

- This experimental result records the maximum run time (in second)

- The randomized sampling costs less computation time than the reservior sampling in processing both two datasets for the randomized sampling traverses graph *G* once, but the reservior sampling needs to visit every node in G twice due to computation of in-degree and out-degree of nodes

Dataset	web-Google	com-LiveJournal
Randomized Sampling	5137.18	1529.02
Reservior Sampling	9255.6	6631.07

Experimental results

- This experiment records two experimental results
  - The maximum numbers of samples
  - The maximum density of samples

- A triangle is considered as the samllest sample by Randomized Sampling, but a node cannot be considered as a cohesive subgraph

Dataset	web-Google	com-LiveJournal
Ran_{max} / Res_{max}	230018 / 13941	8632 / 7039
Ran_{max} / Res_{max}	0.92 / 0.85	0.87 / 0.69

Conclusion

- An approach named randomized sampling algorithm has been proposed for communities extraction in social networks
- The experiment results show that the proposed algorithm is efficient and practicable in various fields
- More experiments will be done to evaluate the application of the proposed algorithm, and observe the memory usage(memory cost) and process speed(time cost)

Q & A

- Please do not hesitate to contact me if you have any questions about the paper
  - yanting8015@hotmail.com
  - yanting8015@sgu.edu.cn

Thank you for your listening.