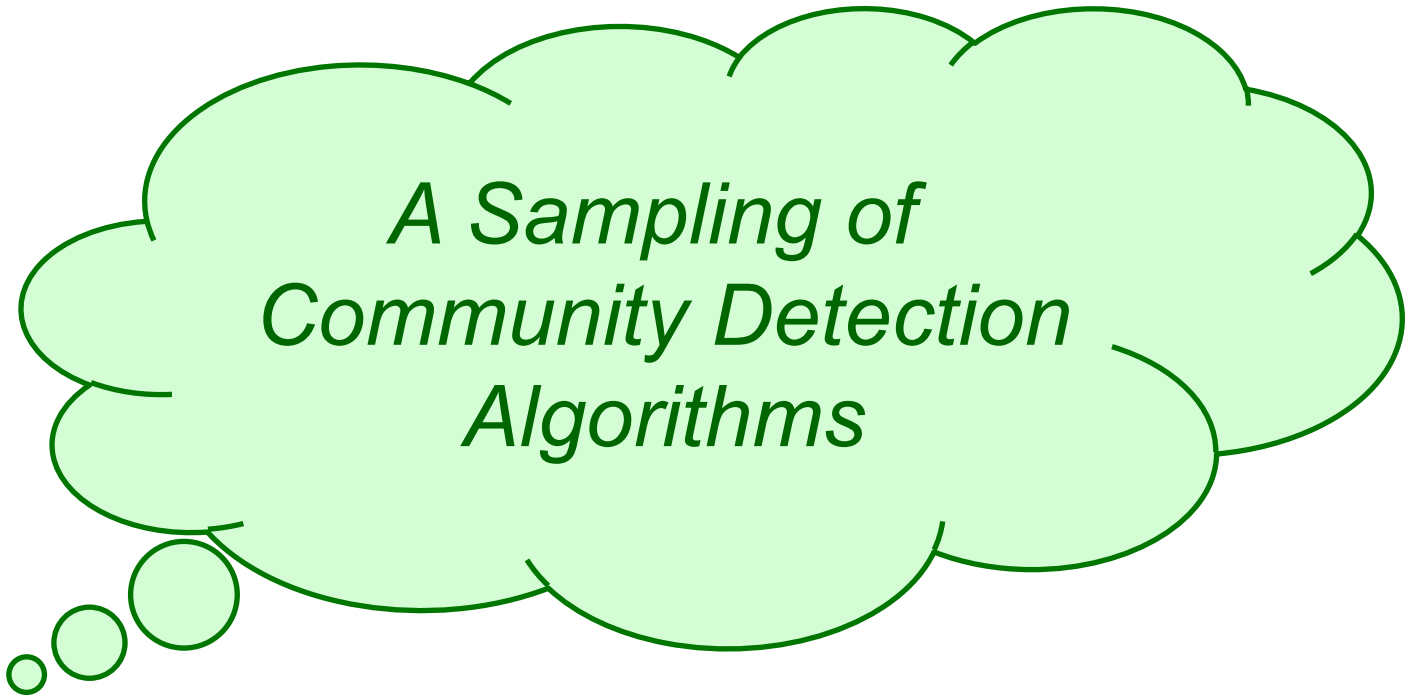
NUS National University of Singapore
School of Computing

Community Discovery Algorithms (An Overview)

Leong Hon Wai (梁汉槐)
School of Computing and USP
National University of Singapore
Email, FB: leonghw@comp.nus.edu.sg

Get an overview of algorithmic community discovery



For more details, go to
Santo Fortunato, “Community detection in Graphs”, Physics Reports, 486 (2010), 75-174.

Good Survey of Algorithms

- Traditional Algorithhms
- Divisive Algorithms
- Modularity-based Methods
- Spectral Algorithms
- Dynamic Algorithms
- Methods for Overlapping Communities
- …

Santo Fortunato, “Community detection in Graphs”, Physics Reports, 486 (2010), 75-174.

Good Survey of Algorithms

- Traditional Algorithhms
- Divisive Algorithms
- Modularity-based Methods
- Spectral Algorithms
  - Min-Cut Partitioning
  - Graph Laplacian
- Dynamic Algorithms…

Santo Fortunato, “Community detection in Graphs”, Physics Reports, 486 (2010), 75-174.

Good Survey of Algorithms

- Traditional Algorithhms
  - Graph Partitioning
  - Hierarchical Clustering (Single-Linkage Clustering)
  - Partitional Clustering
  - Spectral Clustering
- Divisive Algorithms
- . . .

Santo Fortunato, “Community detection in Graphs”, Physics Reports, 486 (2010), 75-174.

Community Structure (informal defn)

"groups of vertices with
***dense intra-group connections*, and**
sparse inter-group connections."

Within-group (intra-group): Very "similar"
(Similar = distance between them is small)

Between-groups (inter-group): Very "dissimilar"
(Dissimilar = distance between them is BIG)

Hon Wai Leong, Computer Science, NUS

From Question T2-D1 (2017)

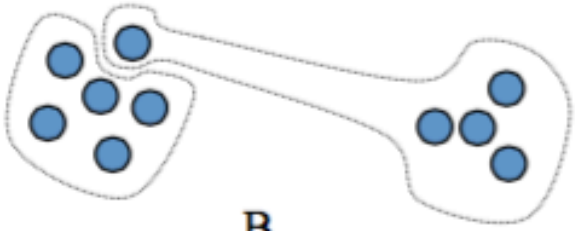
T2-D1: [Simple Community Detection]

In this tutorial question, we will explore a simple community detection problem formulation, that can be surprisingly solved using one of the algorithms covered in the lecture.

You are given n data points and a function d to compute the distance between a pair of data points.

Your goal is to group the **similar** data points into k communities. It also means that the communities should be **dissimilar** with each other. Assume that the number of communities k is known.

A

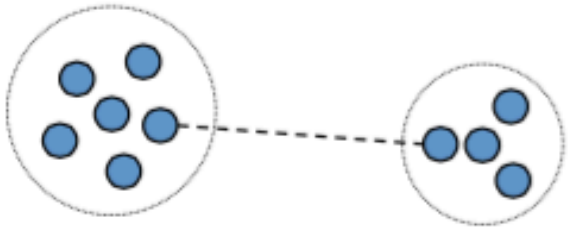


B

We have to define the dissimilarity between cluster so that, for instance in the example above, the communities in A must be scored "better" than the communities in B.

From Question T2-D1 (2017)

One way to define the dissimilarity between the communities is as follows. The data points x and y is called **separated** if they are grouped into two different communities. The **min-distance** between two communities is the closest distance between two separated points in them (illustrated below).



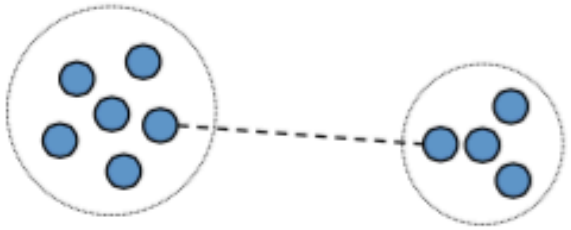
The **spacing** of k communities is the smallest min-distance between all clusters, or equivalently, the distance between the closest pair of separated points. So, the goal is now to find the k communities such that the spacing is maximized. In the example above, the spacing in community A is much larger than the spacing in community B. This problem is known as **single-linked clustering**.

(a) Give a formulation of the **spacing** above.
(The objective function is then to maximize the spacing).

(b) Give an algorithm that can find the communities with the largest spacing.
Hint: Modify an MST algorithm.

From Question T2-D1 (2017)

One way to define the dissimilarity between the communities is as follows. The data points x and y is called **separated** if they are grouped into two different communities. The **min-distance** between two communities is the closest distance between two separated points in them (illustrated below).

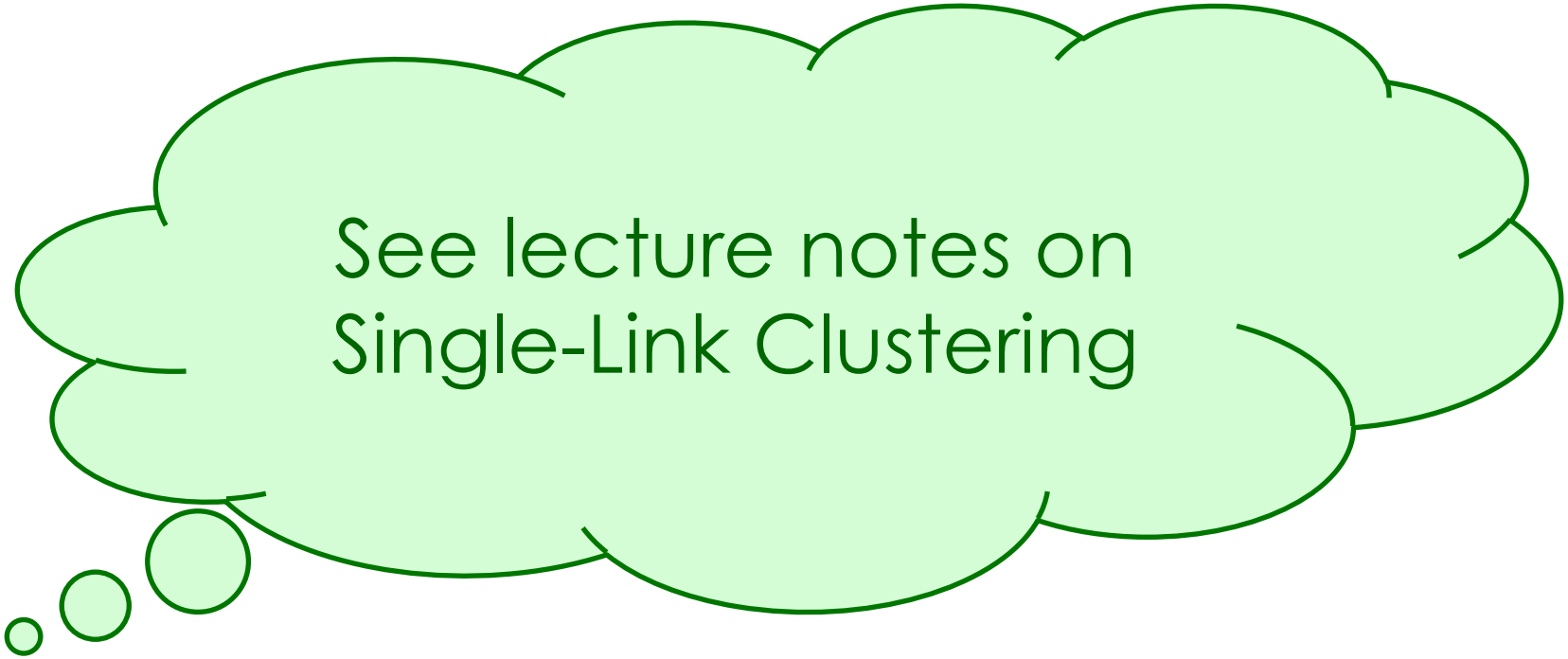


Kruskal's MST Algorithm produces Single Linkage Clustering

The **spacing** of k communities is the smallest min-distance between all clusters, or equivalently, the distance between the closest pair of separated points. So, the goal is now to find the k communities such that the spacing is maximized. In the example above, the spacing in community A is much larger than the spacing in community B. This problem is known as **single-linked clustering**.

(a) Give a formulation of the **spacing** above.
(The objective function is then to maximize the spacing).

(b) Give an algorithm that can find the communities with the largest spacing.
Hint: Modify an MST algorithm.



Good Survey of Algorithms

- Traditional Algorithms
- Divisive Algorithms
  - Girvan and Newman Algorithm
  - Other methods
- Modularity-based Methods
- Spectral Algorithms
- Dynamic Algorithms

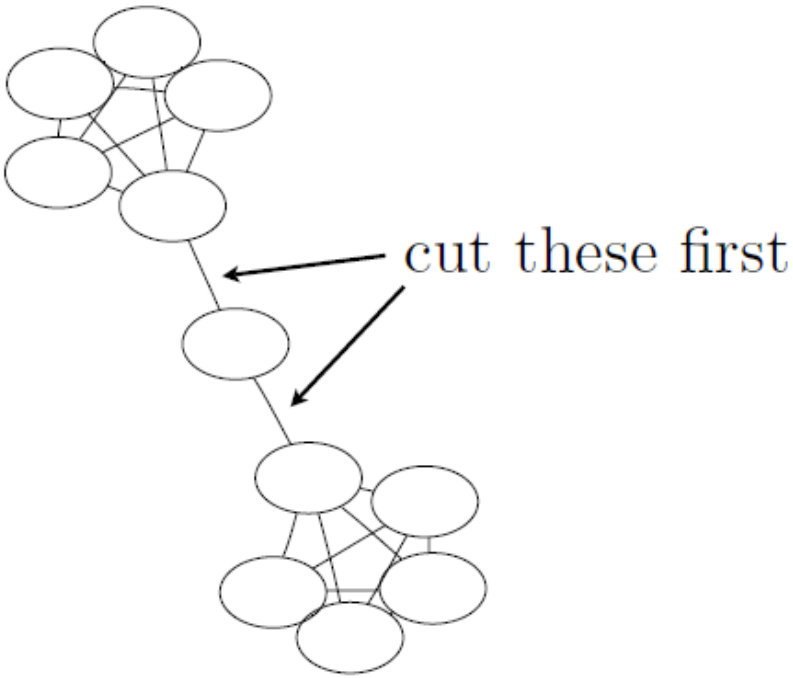
Santo Fortunato, “Community detection in Graphs”, Physics Reports, 486 (2010), 75-174.

Divisive Methods

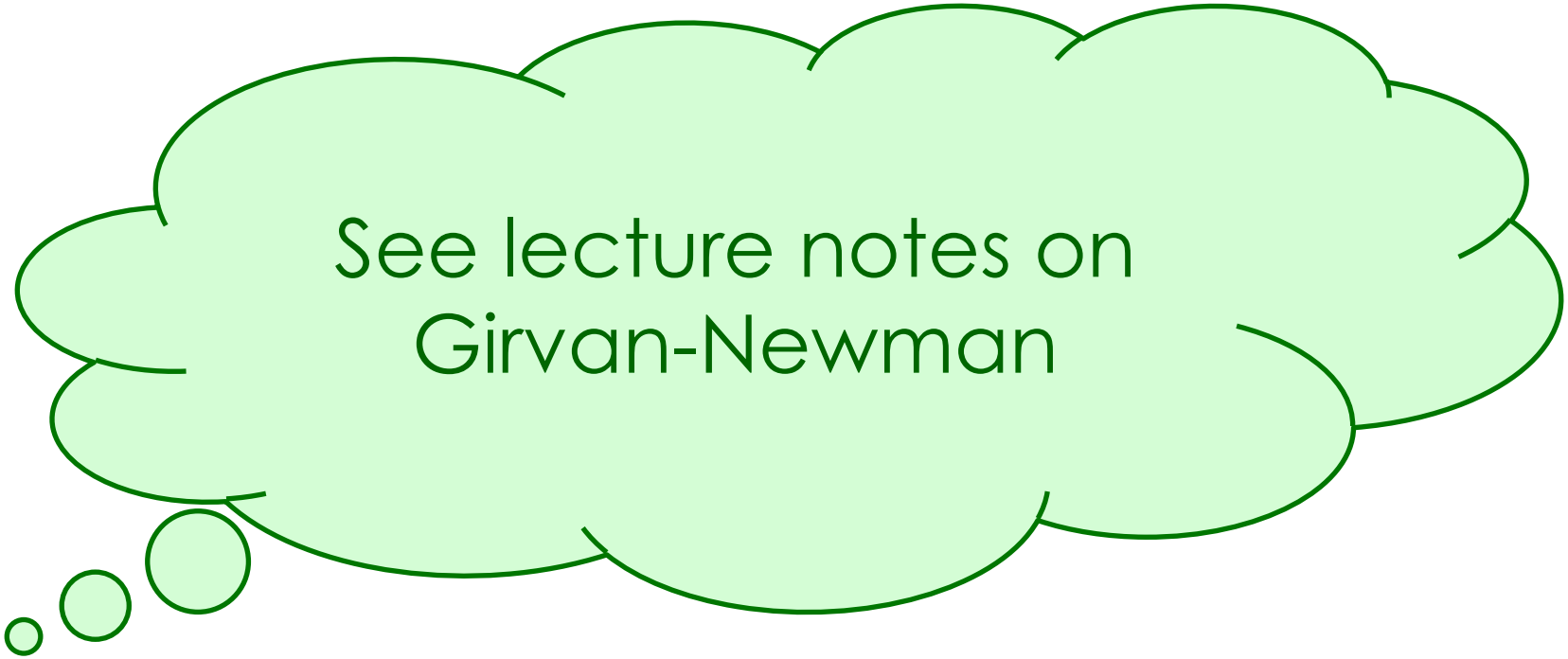
Newman, Girvan. 2002

Prototypical example: **Edge Betweenness**

- betweenness($e_{i,j}$) = number of times $e_{i,j}$ appears in all shortest paths
- High betweenness edges are more "central"



- Expensive, $\mathcal{O}\left(N^3\right)$


See lecture notes on
Girvan-Newman

Good Survey of Algorithms

- . . .
- Methods based on Statistical Inference
- Methods for Overlapping Communities
  - Clique Percolation Method
- Multi-Resolution Methods & Cluster Hierarchy
- . . .

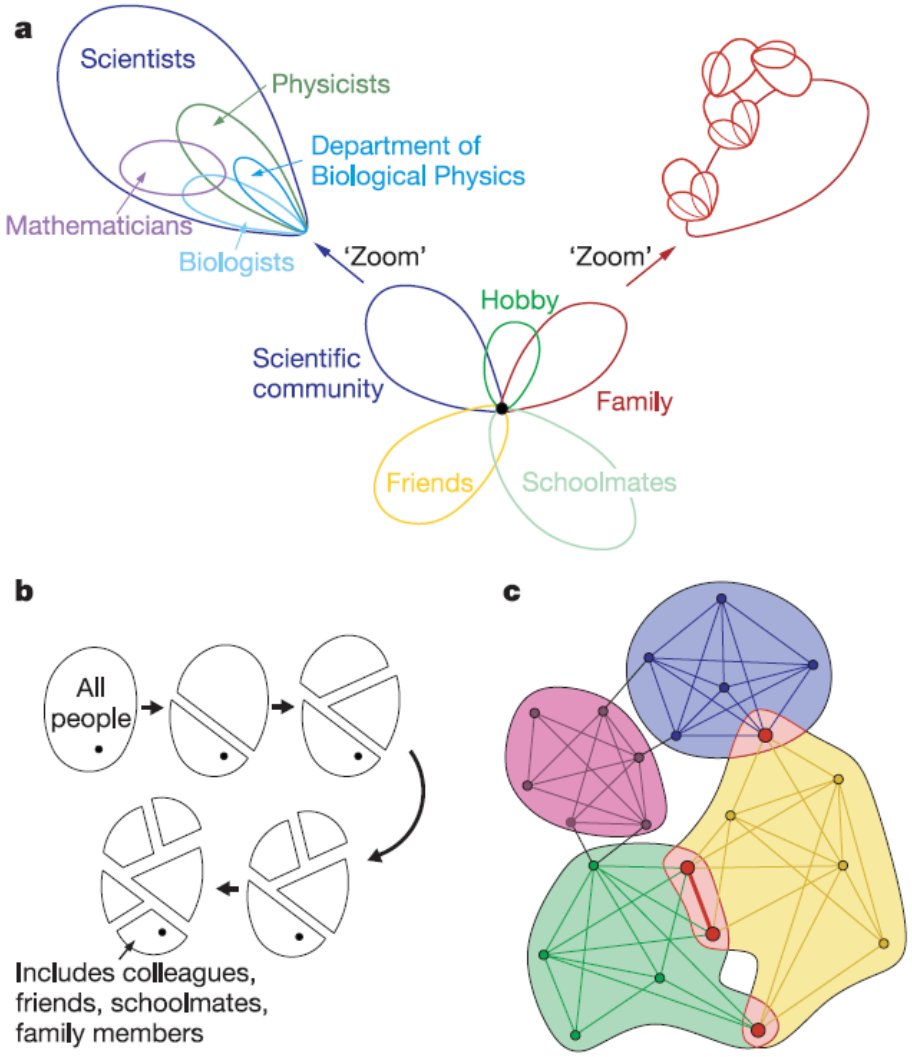
Santo Fortunato, “Community detection in Graphs”, Physics Reports, 486 (2010), 75-174.

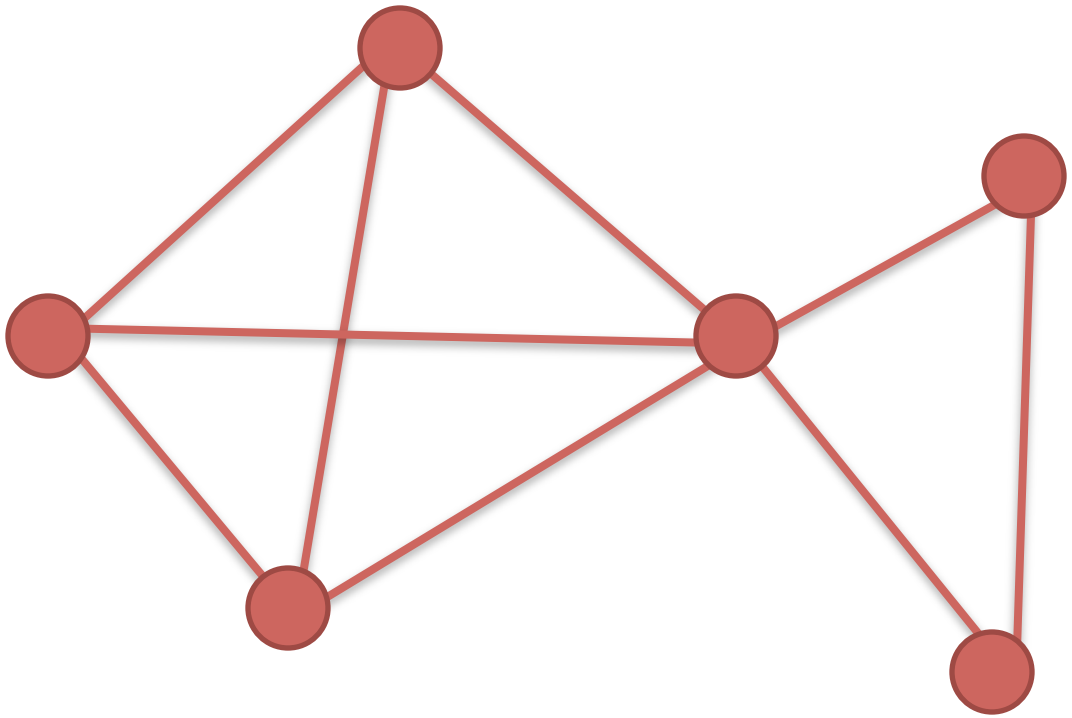
nature Vol 435|9 June 2005|doi:10.1038/nature03607

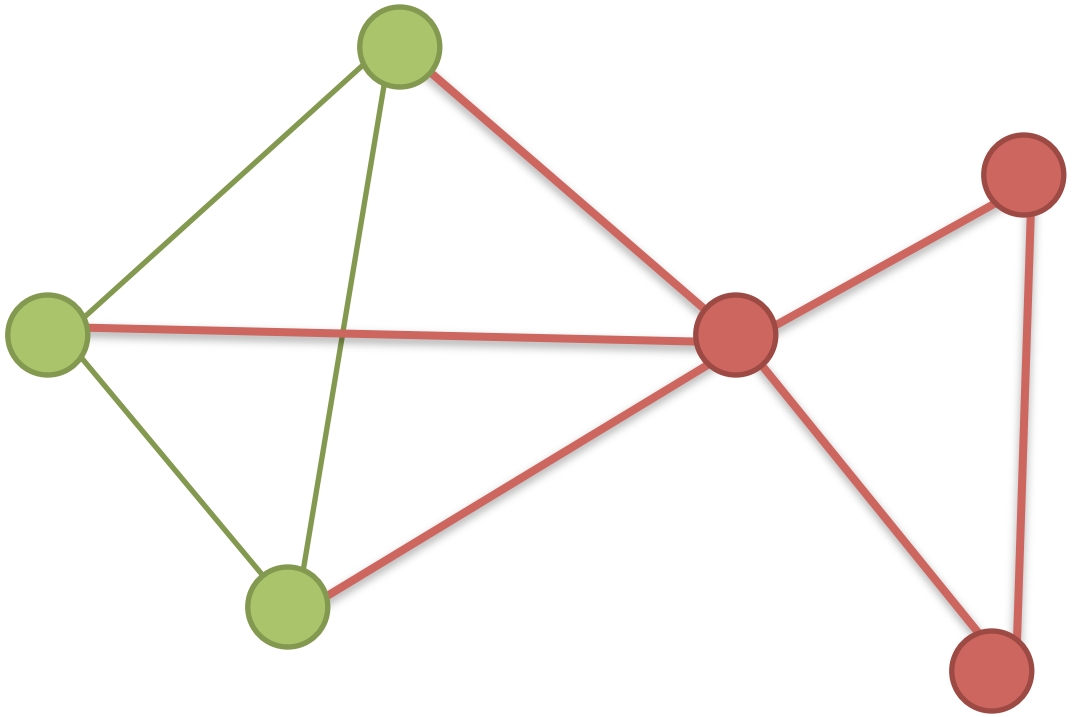
LETTERS

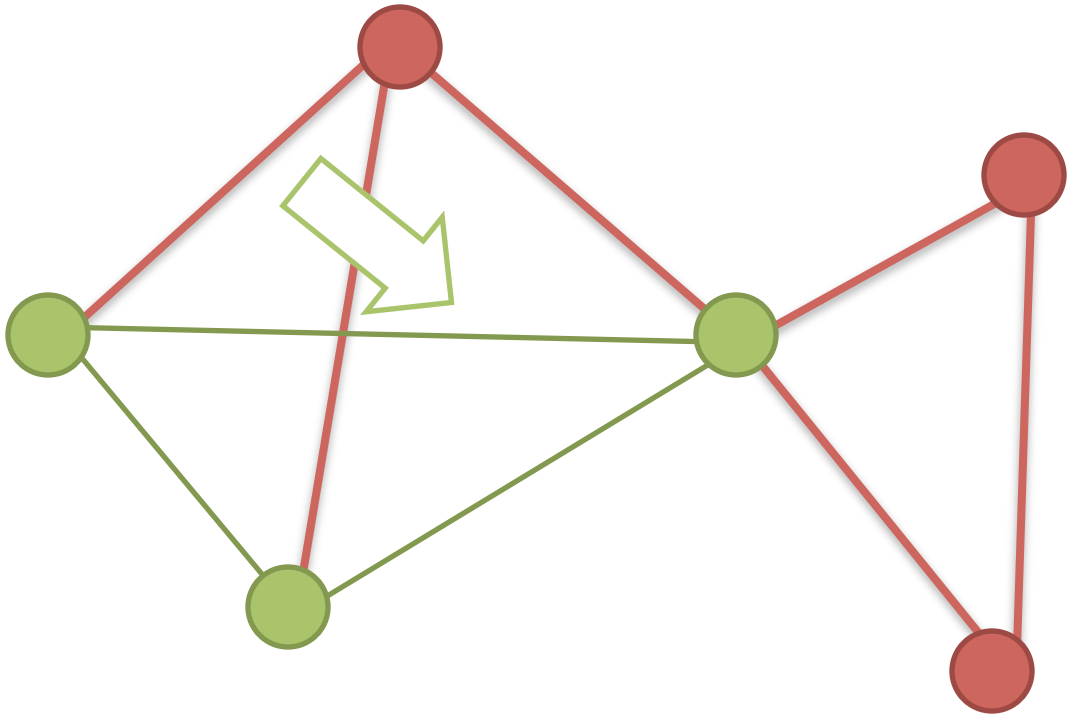
Uncovering the overlapping community structure of complex networks in nature and society

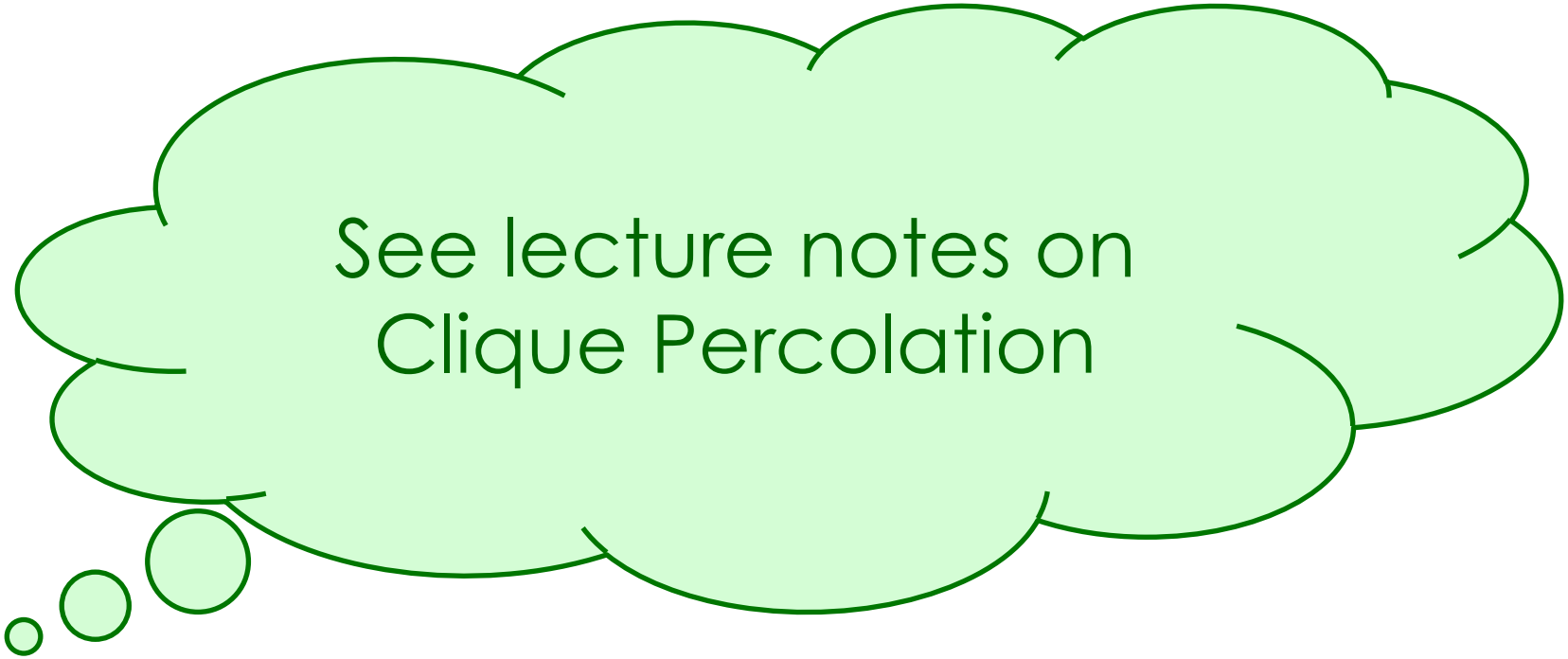
Gergely Palla[1,2], Imre Derényi[2], Illés Farkas[1] & Tamás Vicsek[1,2]









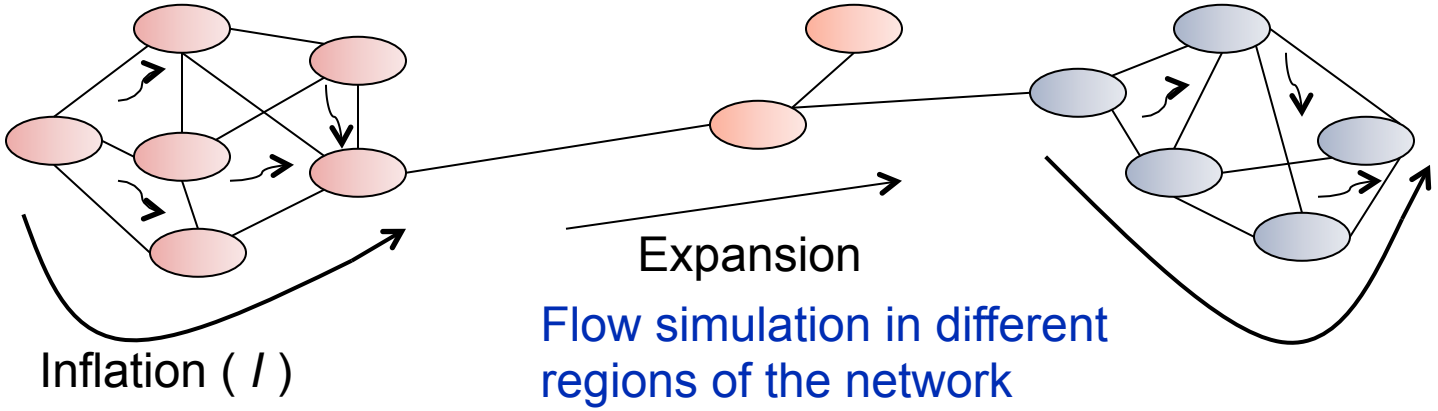

See lecture notes on
Clique Percolation

Good Survey of Algorithms

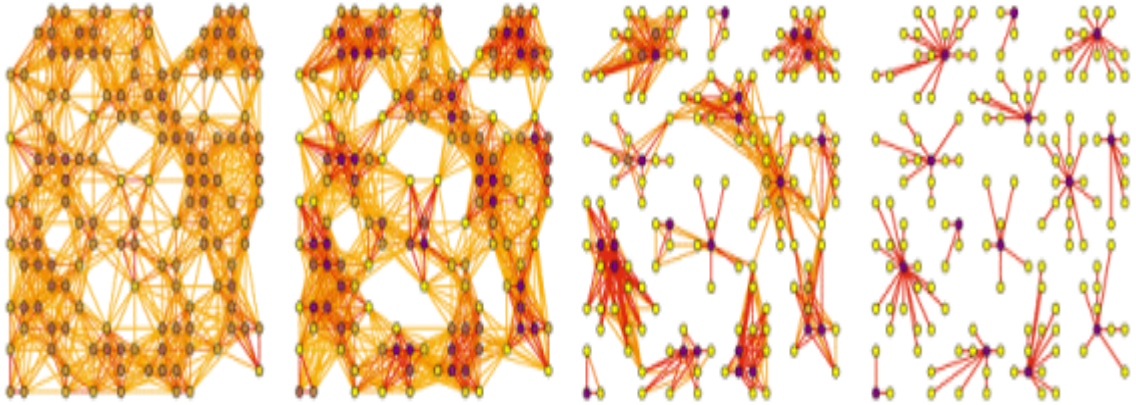
- Traditional Algorithhms
- Divisive Algorithms
- Modularity-based Methods
- …
- Dynamic Algorithms
  - Random Walk, MCL

Santo Fortunato, "Community detection in Graphs", Physics Reports, 486 (2010), 75-174.

MCL (van Dongen, 2000)

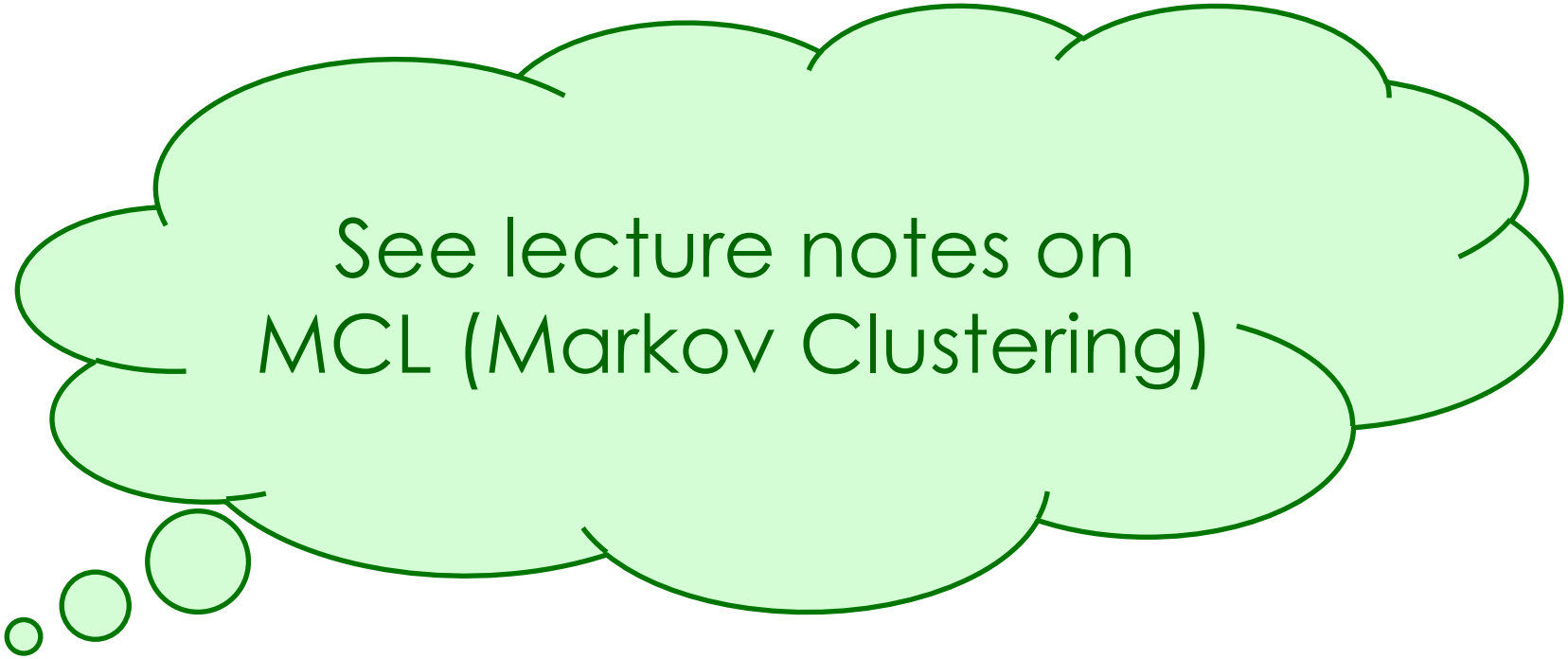


Flow simulation in different regions of the network



Repeated inflation and expansion separates the network into multiple dense regions

Dongen, PhD Thesis, CWI, Netherlands, 2000


See lecture notes on
MCL (Markov Clustering)

Some additional references…

Santo Fortunato, "Community Detection in Graphs", Physics Reports, 486 (2010), 75-174.

Lei Tang, Huan Liu, "*Community Detection and Mining in Social Media*", Morgan and Claypool Publishing, 2010.

http://dmml.asu.edu/cdm/

http://www.cscs.umich.edu/~crshalizi/notebooks/community-discovery.html

NUS National University of Singapore

School of Computing

Thank you.

Q & A

Contact:

Hon Wai Leong（梁汉槐）

FB, email: leonghw@comp.nus.edu.sg

http://www.comp.nus.edu.sg/~leonghw/