



# Mining Communities in Large Networks from Big Data

Leong Hon Wai (梁汉槐)

Department of Computer Science  
National University of Singapore

[leonghw@comp.nus.edu.sg](mailto:leonghw@comp.nus.edu.sg)  
<http://www.comp.nus.edu.sg/~leonghw/>



Talk @ MST, Fri, Nov 06, 2015

# Outline of Talk

---

- History of this Course
- Big Data is Everywhere
- Big Data gives Large Networks
- Community Structures
- Mining Community in Large Networks
- Demo (from a recent undergraduate project)
- Algorithms for Community Detection



# Our research in CD in Networks (1)

## Started with Finding Hubs and Quasi Cliques

- ❑ S. Srihari\*, H.K. Ng, K. Ning, H.W. Leong, "Detecting Hubs and Quasi Cliques in Scale-free Networks," *ICPR*, Tampa, (2008).

## Finding Hubs and Essential Genes in PPI Networks

- ❑ K. Ning, H.K. Ng, S. Srihari, H.W. Leong, and A.I. Nesvizhskii, "Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology," *BMC Bioinformatics* Vol. 11:505 (Oct 2010).

# Our research in CD in Networks (2)

## Predicting Communities (Protein Complexes) in PPI Networks

- ❑ S. Srihari, K. Ning, and H.W. Leong, "Refining Markov Clustering for Protein Complex Prediction by Incorporating **Core-attachment Structure**," *Genome Informatics*, Vol. 23, (2009), pp. 159-166.

## Community Detection in **Weighted** PPI Networks

- ❑ S. Srihari, and H.W. Leong, "MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure," *BMC Bioinformatics*, Vol. 11:504 (Oct 2010).

# Our research in CD in Networks (3)

## Dealing with *Missing Data* and *Sparse Complexes*

- ❑ S. Srihari, and H.W. Leong, "Employing functional interactions for characterisation and detection of sparse complexes from yeast PPI networks", IJBRA, 8:3/4, 286-304, (2012). .

## Protein Complexes are *Dynamic (Time Dependent)*

- ❑ S. Srihari, K. Ning, and H.W. Leong, "Temporal Dynamics of Protein Complexes in PPI Networks: A Case Study using Yeast Cell Cycle Dynamics," BMC Bioinformatics, Dec 2012.

## Leveraging conservation of interactions...

- ❑ PV Nguyen, S. Srihari and HWi Leong, "Identifying **conserved protein complexes** between species by constructing interolog networks," BMC Bioinformatics, 2013, 14(S-16):S8

# Our research in CD in Networks (4)

## Put together a Tutorial on Complex Detection from PPI

- ❑ S. Srihari and H.W. Leong, "A Tutorial on Computational Methods for Protein Complex Prediction from Protein Interaction Networks," INCoB, (Oct 2012) and elsewhere since.

## Writing a Survey Paper...

- ❑ Sriganesh Srihari and Hon Wai Leong, "A Survey of Computational Methods for Protein Complex Prediction from Protein Interaction Networks," JBCB, (2013), 11:2, (2013)

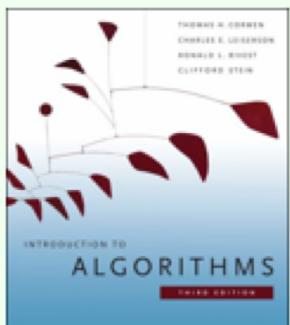
# Our research in CD in Networks (5)

## Turning Research into Undergraduate Education (Sp 2014)

- ❑ 8 undergraduates working on CD in Large Networks
- ❑ A student, Yujian YAO (sophomore) developed a Cool FB App

## Turning this into a Summer Course (Sp 2016)

- ❑ Lecture at UTAR Big-Data Workshop (May 2014)
- ❑ Creating a summer course for SWS3001 for NUS



## CS3230R: Design and Analysis of Algorithms (R)

Spring 2014

R Section for CS3230R of Fall 2013, Spring 2014  
(Leong Hon Wai and Ken Sung Wing Kin)

Course Web-Site

### LATEST UPDATE

- **27-Mar-2014 (W10): Talk by Yu ShuZhi, Nguyen Truong Huy**  
Links to the Data-Sets for Comm Detection is coming up soon.
- **20-Mar-2014 (W9): Talk by YAO YuJian and Darius Foo**
  - ShuZhi to chair the session.
  - Prof Leong away on conference in Japan -- | [NII-Shonan-Meeting #45](#)  
"Towards the Ground Truths: Exact Algorithms for Bioinformatics Research"

(See bottom of page for past announcements...)

(Mining Communities) Page 9

# Trip to Prof Kal's university (May'14)



## Detecting Communities in Large Social Networks

Leong Hon Wai (梁汉槐)

Department of Computer Science  
National University of Singapore



leonghw@comp.nus.edu.sg  
<http://www.comp.nus.edu.sg/~leonghw/>

Talk @ UTAR, 8 May 2014

Hon Wai Leong, NUS

(Detecting Communities) Page 1

© Leong Hon Wai, 2014

<https://www.facebook.com/photo.php?fbid=10152122148331239&set=a.10152116054231239&type=3>

Hon Wai Leong, Computer Sci

© Leong Hon Wai, 2015



# BIG data



NYT, Feb 11, 2012: “The Age of Big Data”

- “What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. ...”

A lot of people are talking about big data, but most people are just creating it

The real value is in the analysis

And all others (we) are indeed bringing data!

# What happens in an Internet minute!

Every two days now we create as much information as we did from the dawn of civilization up until 2003 (Eric Schmidt, Google CEO)





# Big Data and Large networks

---

## □ Big Data is notable,

- ❖ not just because of its size,
- ❖ but because of its relationship to other data.

## □ Big Data

- ❖ Are collected & aggregated
- ❖ Are fundamentally networked
- ❖ Data are threaded with connections

# Big Data and Large networks

---

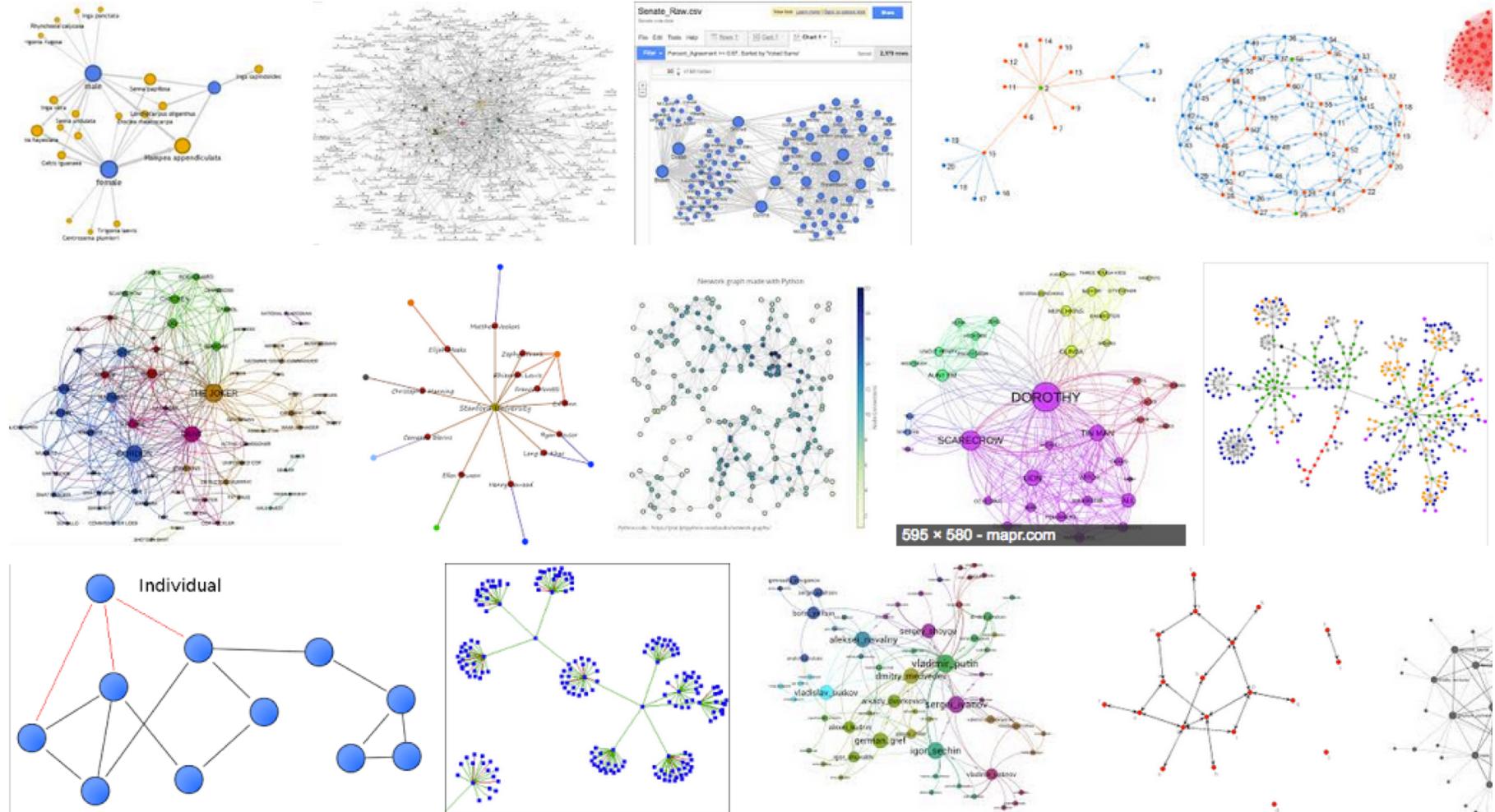
## □ The value of big data comes from

- ❖ the patterns that can be derived by making connections between pieces of data,

## □ The Pattern could be

- ❖ About an individual,
- ❖ Or about individuals in relation to others,
- ❖ Or about groups of people,
- ❖ Or simply about the structure of information.

# Large Networks are A&E



(Mining Communities) Page 17

# Large Real-World Networks

---

- Internet graphs, WWW graphs
- Citation networks, actor networks
- Transportation network, Email networks
- Food Web,
- Social Networks (FB, Linked-In, etc)
- Biochemical networks
- **Protein-Protein Interaction (PPI) networks**

# Networks from Big-Data

---

## □ Nodes

People, Customers, household/family, patient, doctor, author, terrorist, webpage, Items

## □ Edges (directed or undirected)

- ❖ Relationship of some kind,  
e.g. FB-friends, colleagues, disease, contact, reference, bought-together
- ❖ (could be weighted based on importance, frequency, reliability, etc)

# Networks from BigData: Examples

---

## □ Classic Supermarket Example

- ❖ Nodes are items
- ❖ Edge indicates the two items in same receipt

## □ Classic Terrorist Network

- ❖ Nodes are people
- ❖ Edge if there is some contact between them

## □ Churn detection in Telco

- ❖ Nodes are customers
- ❖ Edges are calling patterns between customers

# Networks from BigData: Examples

---

## ❑ PPIN (protein-protein interaction network)

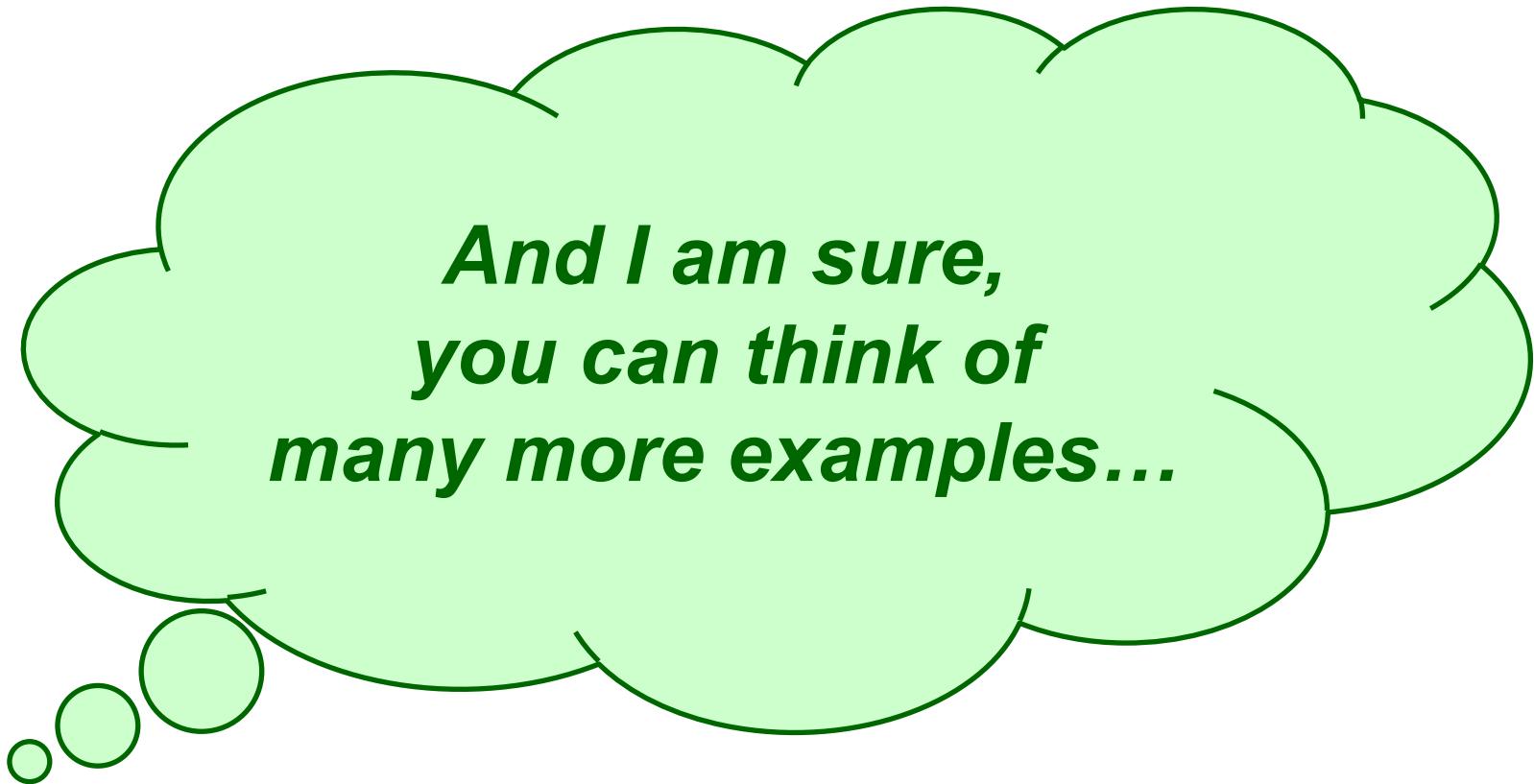
- ❖ Nodes are proteins
- ❖ Edge indicates the two proteins interact

## ❑ Anti-Money Laundering

- ❖ Nodes are bank account
- ❖ Edges are money transfer

## ❑ System Risk in a Credit Risk setting

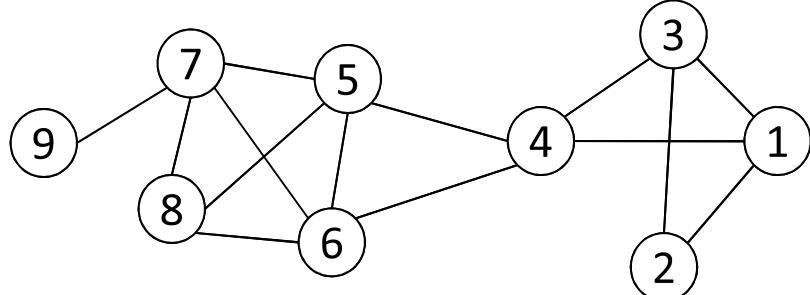
- ❖ Nodes are banks
- ❖ Edges are liquidity dependencies



# Networks and Representation

**Network:** A graph structure made of nodes (individuals or organizations) and edges that connect nodes in various relationships like friendship, kinship etc.

- Graph Representation
- Matrix Representation

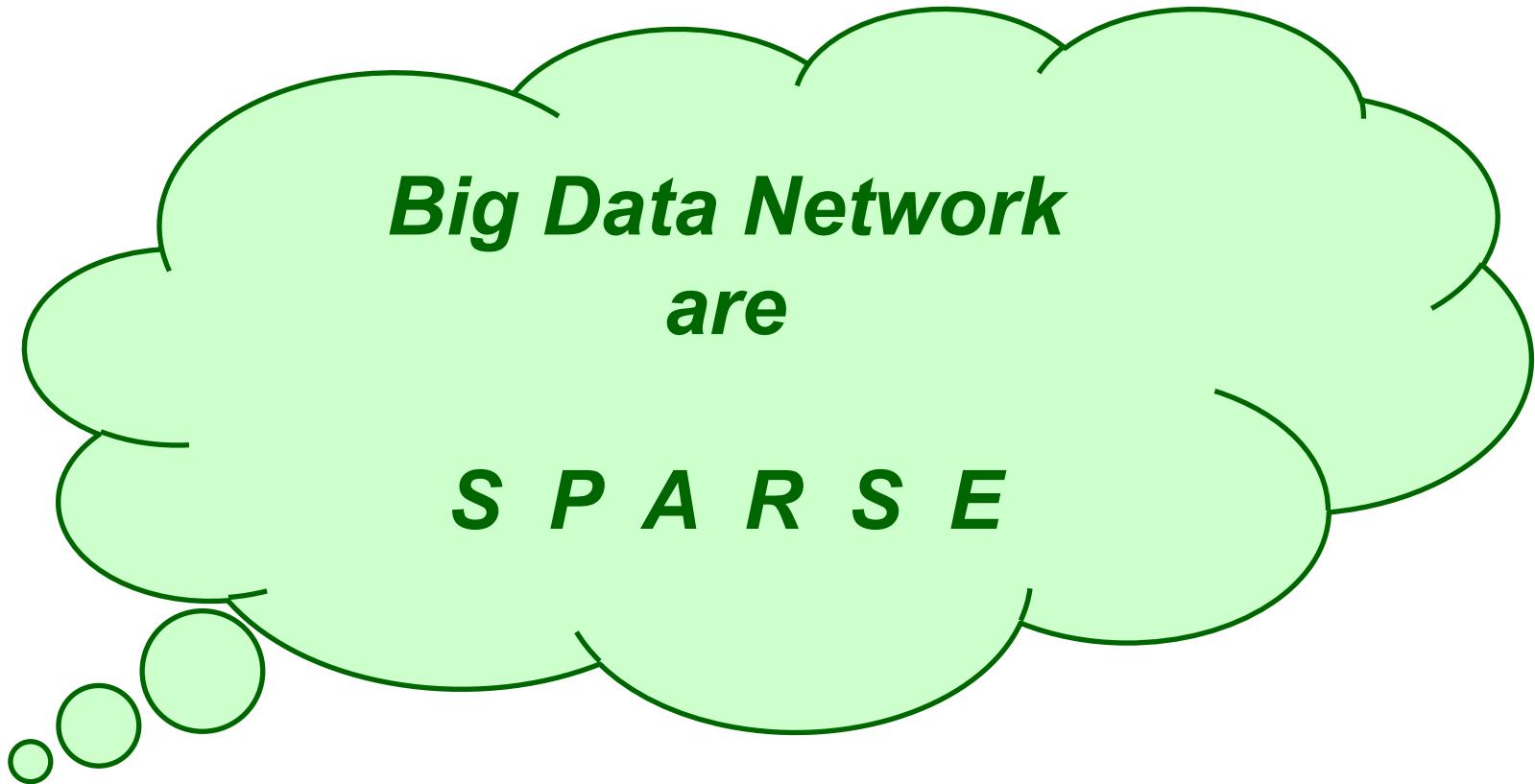


| Node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------|---|---|---|---|---|---|---|---|---|
| 1    | - | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2    | 1 | - | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3    | 1 | 1 | - | 1 | 0 | 0 | 0 | 0 | 0 |
| 4    | 1 | 0 | 1 | - | 1 | 1 | 0 | 0 | 0 |
| 5    | 0 | 0 | 0 | 1 | - | 1 | 1 | 1 | 0 |
| 6    | 0 | 0 | 0 | 1 | 1 | - | 1 | 1 | 0 |
| 7    | 0 | 0 | 0 | 0 | 1 | 1 | - | 1 | 1 |
| 8    | 0 | 0 | 0 | 0 | 1 | 1 | 1 | - | 0 |
| 9    | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | - |

# Outline of Talk

---

- Big Data is Everywhere
- Big Data gives Large Networks
- □ Community Structures
- Demo (from a recent undergraduate project)
- Detecting Community in Large Networks
- Application Example  
*(from Computational Biology)*

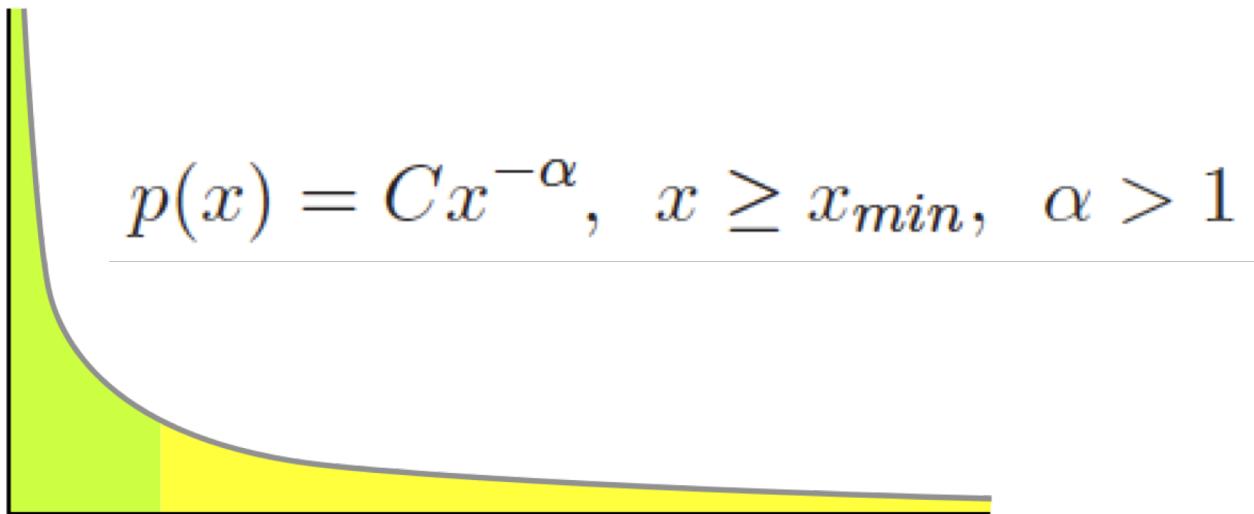


# Properties of Large-Scale Networks

- Networks from big-data are typically huge, involving millions of nodes and connections.
- Large-scale networks in real world demonstrate similar patterns
  - Scale-free distributions
  - Small-world effect
  - Strong Community Structure

# Scale-free Distributions

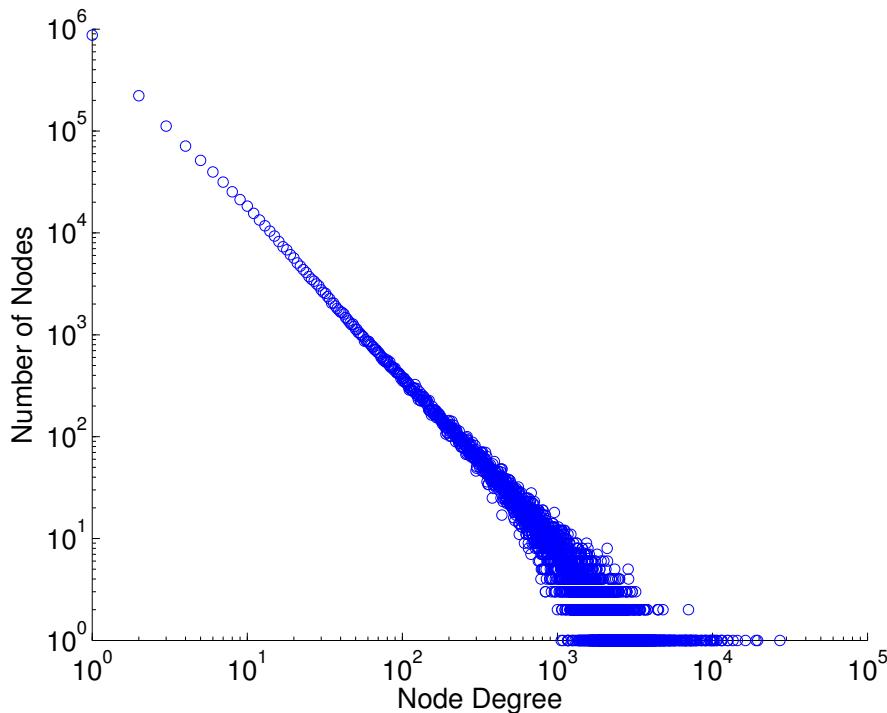
- Degree distribution in large-scale networks often follows a **power law**.



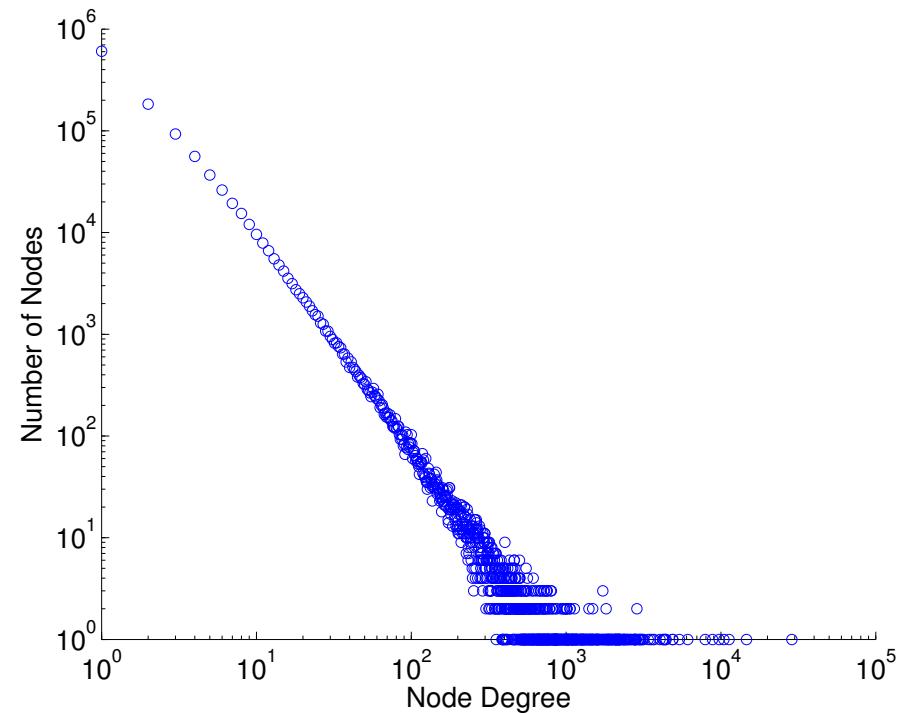
- A.k.a. **long tail** distribution, **scale-free** distribution

# log-log plot

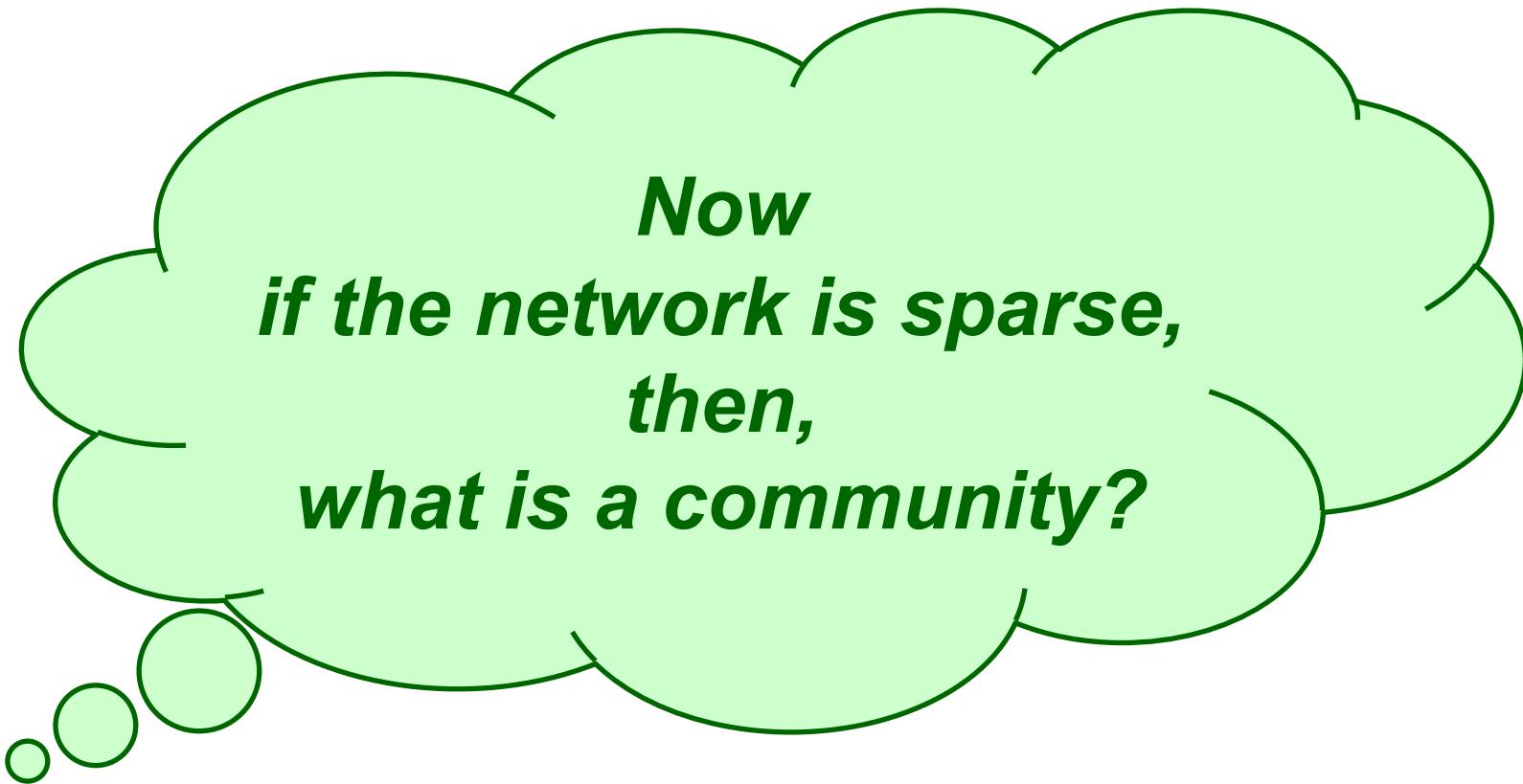
- Power law distribution becomes a **straight line** if plot in a log-log scale



Friendship Network in Flickr



Friendship Network in YouTube



# From St. Louis to Rolla

---

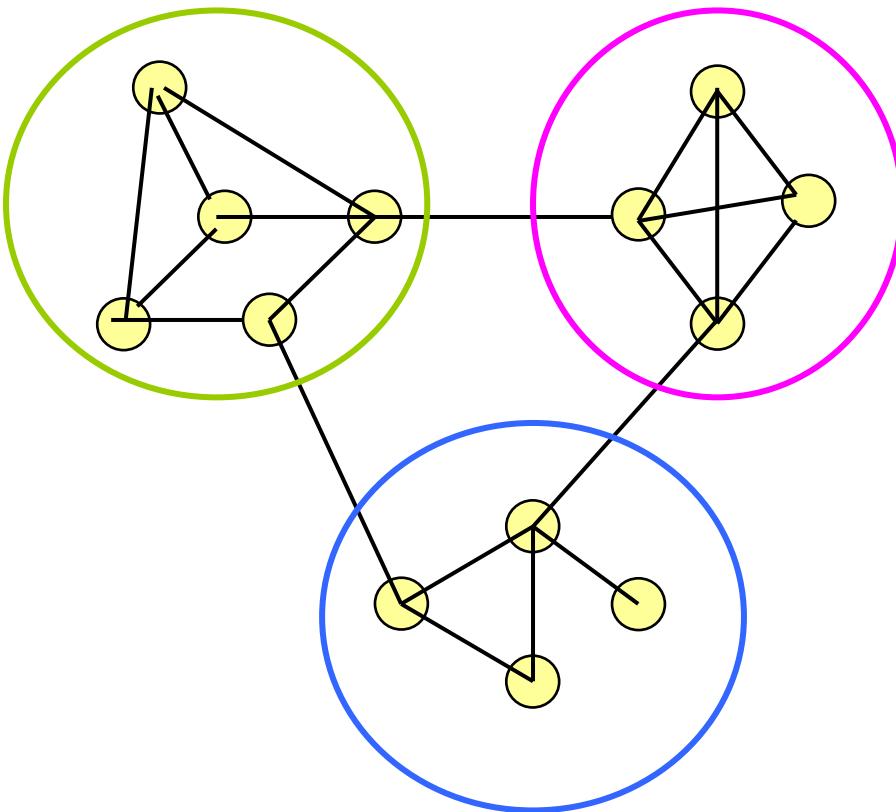
- ❑ My drive from St. Louis to Rolla:

<https://bit.ly/2Xop3yv>

- ❑ First see the overall map (I57, I70, I55, I44)
- ❑ Zoom in on St-Clair
- ❑ Zoom in on Sullivan
- ❑ Zoom in on Cuba
- ❑ Zoom in on St James
- ❑ Zoom in on Rolla, Missouri

Illustration by  
Road Maps

# Community Structure (example)



# Density of a Graph

---

## □ If a graph $G=(V,E)$ with $n$ nodes, $e$ edges

- ❖ Density =  $e / (n(n-1)/2)$
- ❖ A clique  $K_n$  has density = 1,
- ❖ An empty graph ( $E=\emptyset$ ) has density = 0,
- ❖ A spanning-tree  $T$  has density =  $2/n$ .

# Community Structure (informal defn)

---

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

**Within-group (intra-group) edges:** High density

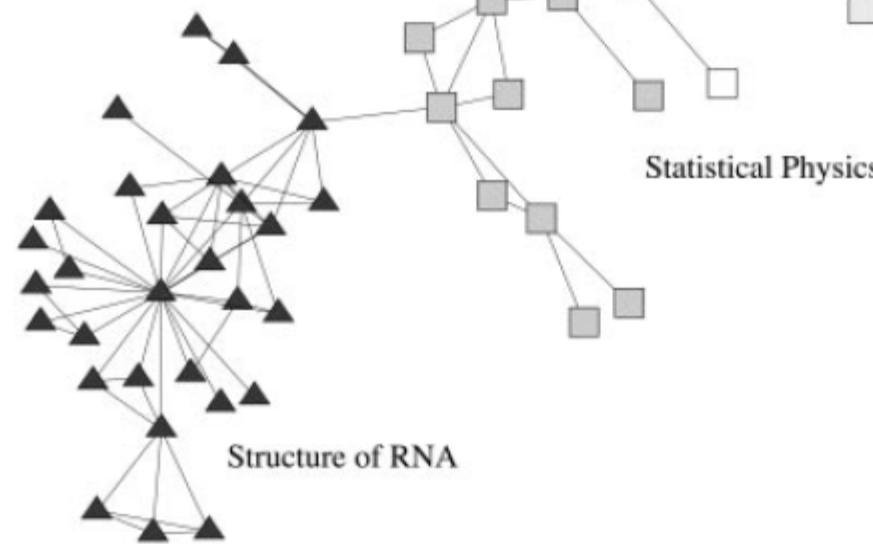
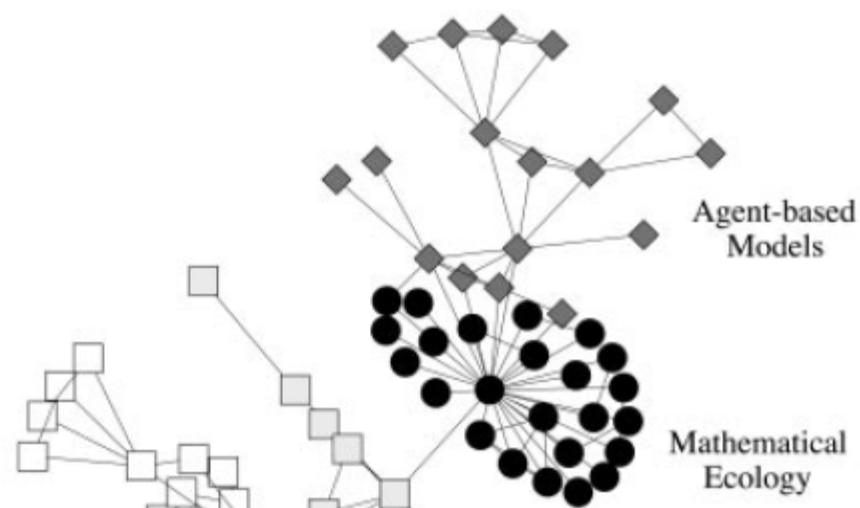
**Between-group (inter-group) edges:** Low density

# Examples of Community Structures

---

- Communities of biochemical network might correspond to “functional units” of some kind.
  
- Communities of a web graph might correspond to sets of “web sites dealing with a related topics”.

## Different Communities among Collaborations of SFI (Santa Fe Institute)

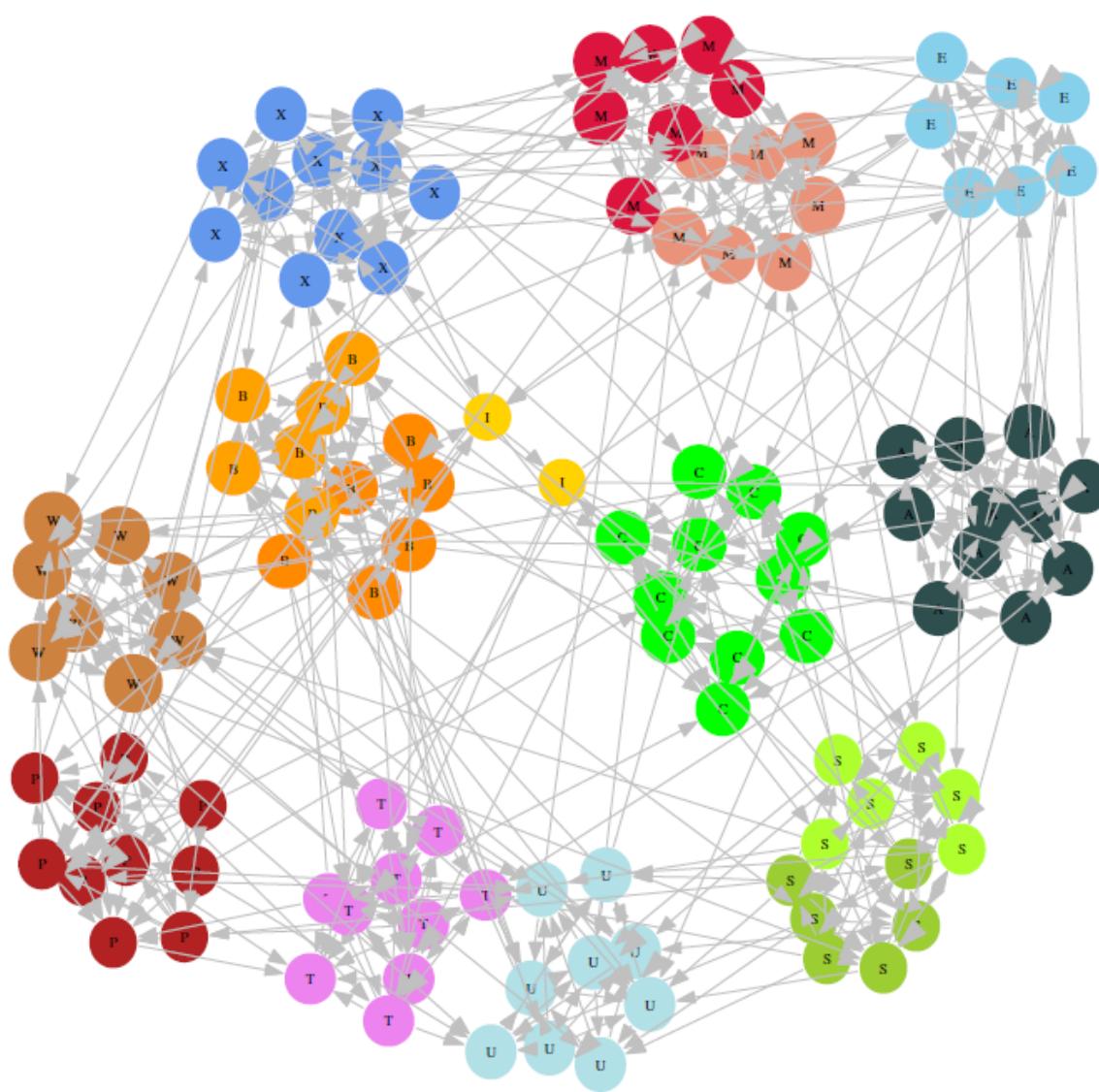


Community structure in social and  
biological networks

M. Girvan<sup>\*†‡</sup> and M. E. J. Newman<sup>\*§</sup>

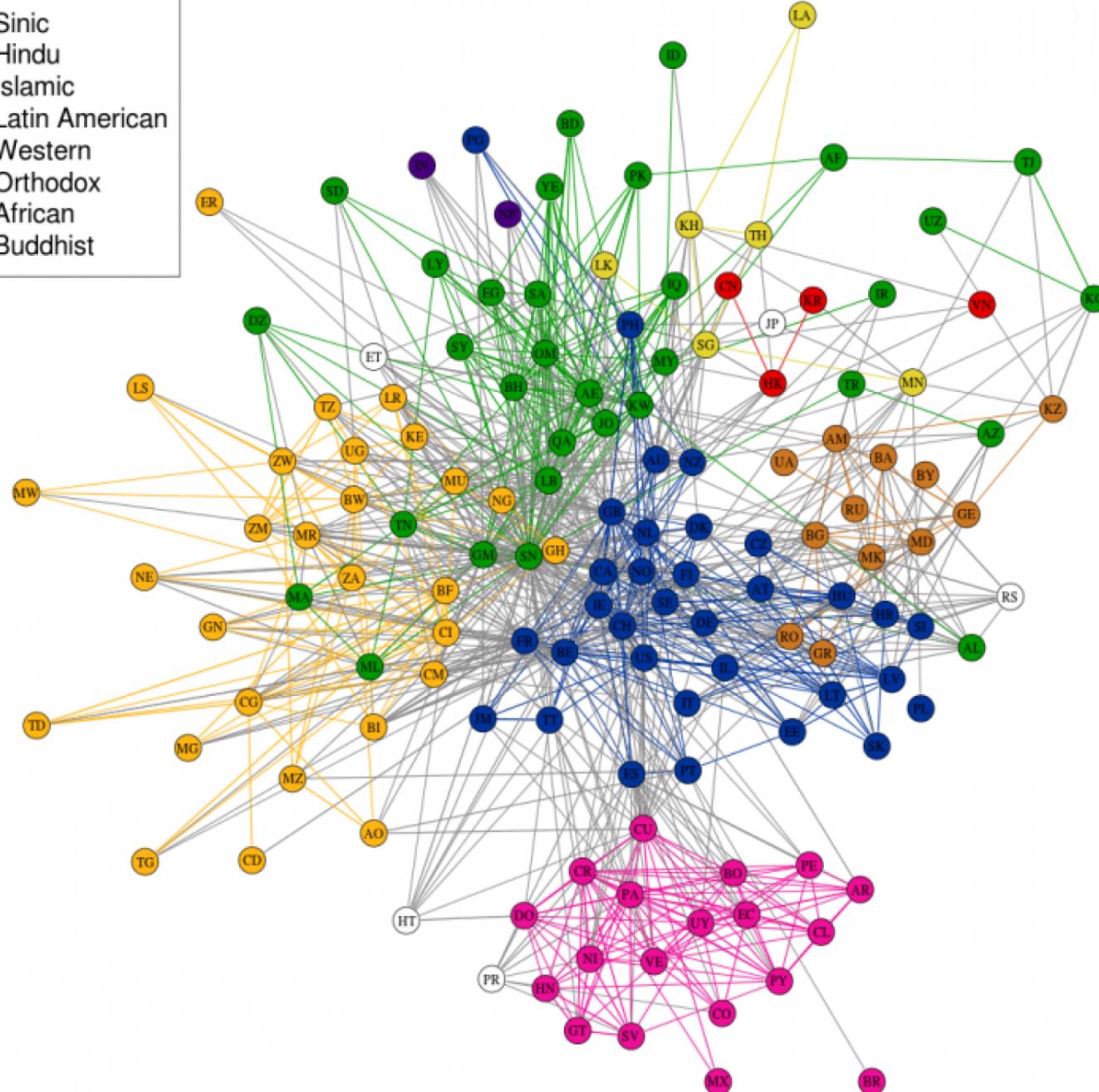
Output of  
Girvan-Newman Alg

# *US College Football Conferences*



**J**ournal of Statistical Mechanics: Theory and Experiment  
An IOP and SISSA journal

**A network-based ranking system for US college football**





***So, what can  
communities  
be good for?***

# Dilemma in Big Data analysis

---

❑ Know what we know

❑ Know what we don't know

❑ Don't know what we know

❑ Don't know what we don't know

*What can we do?  
How can CD help?*

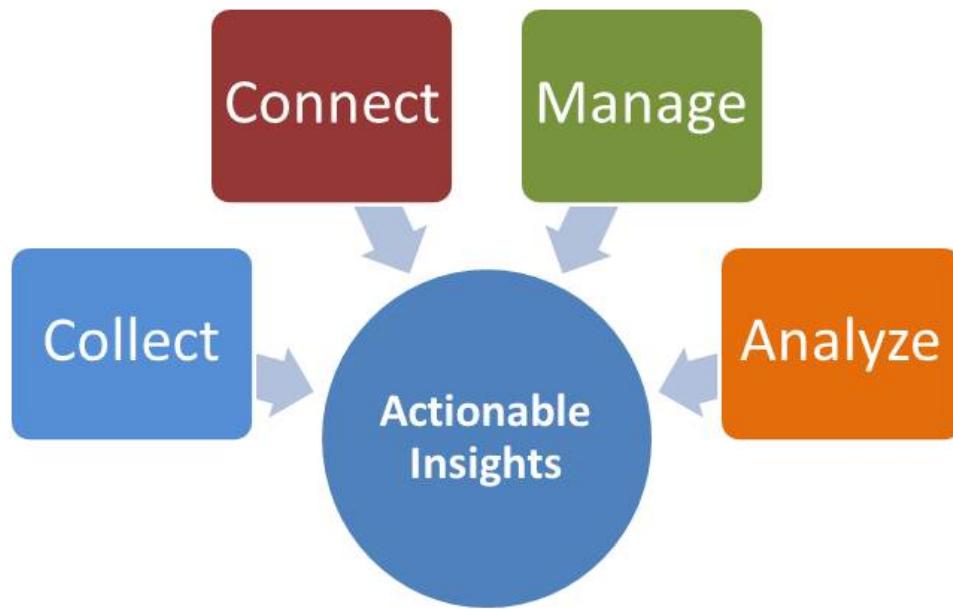
# Use of Community Detection (CD)

---

1. Big data gives us big graphs
2. Detect communities in these graphs
3. Analyze the communities detected
  - ❖ Derive/confirm known relationships / facts

- ❖ Derive new insights  
(betw. community & relationships)
  - ❖ Can turn into competitive advantage?

# New buzzword...



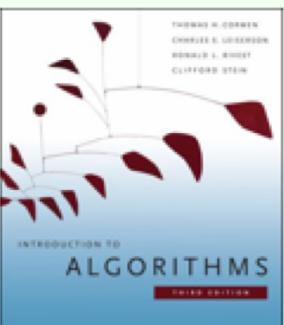
# A Cool project...

---



*How \*well\* do you  
know the communities  
in your FB friends?*

<http://pals.yjyao.com>



## CS3230R: Design and Analysis of Algorithms (R) Spring 2014

R Section for CS3230R of Fall 2013, Spring 2014  
(Leong Hon Wai and Ken Sung Wing Kin)

Course Web-Site

### LATEST UPDATE

- **27-Mar-2014 (W10): Talk by Yu ShuZhi, Nguyen Truong Huy**  
Links to the Data-Sets for Comm Detection is coming up soon.
- **20-Mar-2014 (W9): Talk by YAO YuJian and Darius Foo**
  - ShuZhi to chair the session.
  - Prof Leong away on conference in Japan -- | [NII-Shonan-Meeting #45](#)  
"Towards the Ground Truths: Exact Algorithms for Bioinformatics Research"

(See bottom of page for past announcements...)

(Mining Communities) Page 43

# An Undergraduate Student Project

---

**CS3230R:**

A research-component add-on to  
CS3230: Analysis of Algorithm

**Yao YuJian:** (2<sup>nd</sup> year undergrad, NUS-SOC)

“Improved Fuzzy Clustering and  
Force-Directed Algorithms for  
Visualization of Personal Social Networks”

# YuJian's Project Overview...

---

- ❑ Download your FB friend network
  
- ❑ Compute communities
  - ❖ Using an Improved Fuzzy Clustering Algorithm
  
- ❑ Visualize Communities computed
  - ❖ Using Force-Directed Layout

<http://pals.yjyao.com>

http://pals.yjyao.com

Home

pals

# pals

Visualize your social network!

Graph with names

Graph without names

2014 version  
(offline now)

Revived by Dr. Kal  
(2016--now)

# Trip to Prof Kal's university (May'14)



## Detecting Communities in Large Social Networks

Leong Hon Wai (梁汉槐)

Department of Computer Science  
National University of Singapore



leonghw@comp.nus.edu.sg  
<http://www.comp.nus.edu.sg/~leonghw/>

Talk @ UTAR, 8 May 2014

Hon Wai Leong, NUS

(Detecting Communities) Page 1

© Leong Hon Wai, 2014

<https://www.facebook.com/photo.php?fbid=10152122148331239&set=a.10152116054231239&type=3>

Hon Wai Leong, Computer Sci

© Leong Hon Wai, 2015



(Mining Communities) Page 48

## Outline of Talk

- Big Data is Everywhere
- Big Data gives Large Networks
- Community Structures
- Detecting Community in Large Networks
- Demo (from a recent undergraduate project)
- Application Example  
*(from Computational Biology)*

CPCC Complex Diseases, Sep 2013 | Page 2

Hon Wai Leong, Computer Science, NUS

© Leong Hon Wai





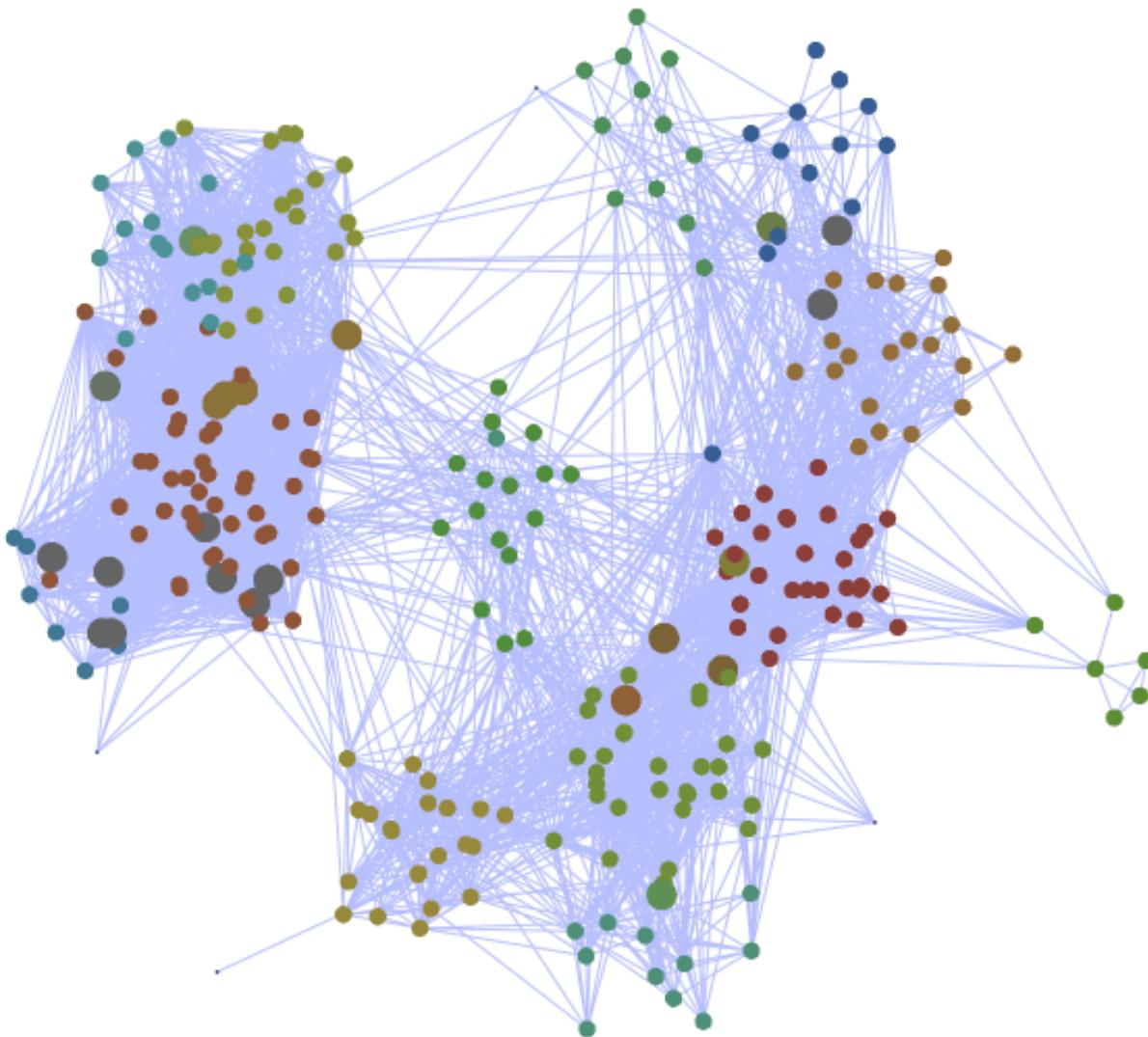
- 
- Try out the app written by YuJian

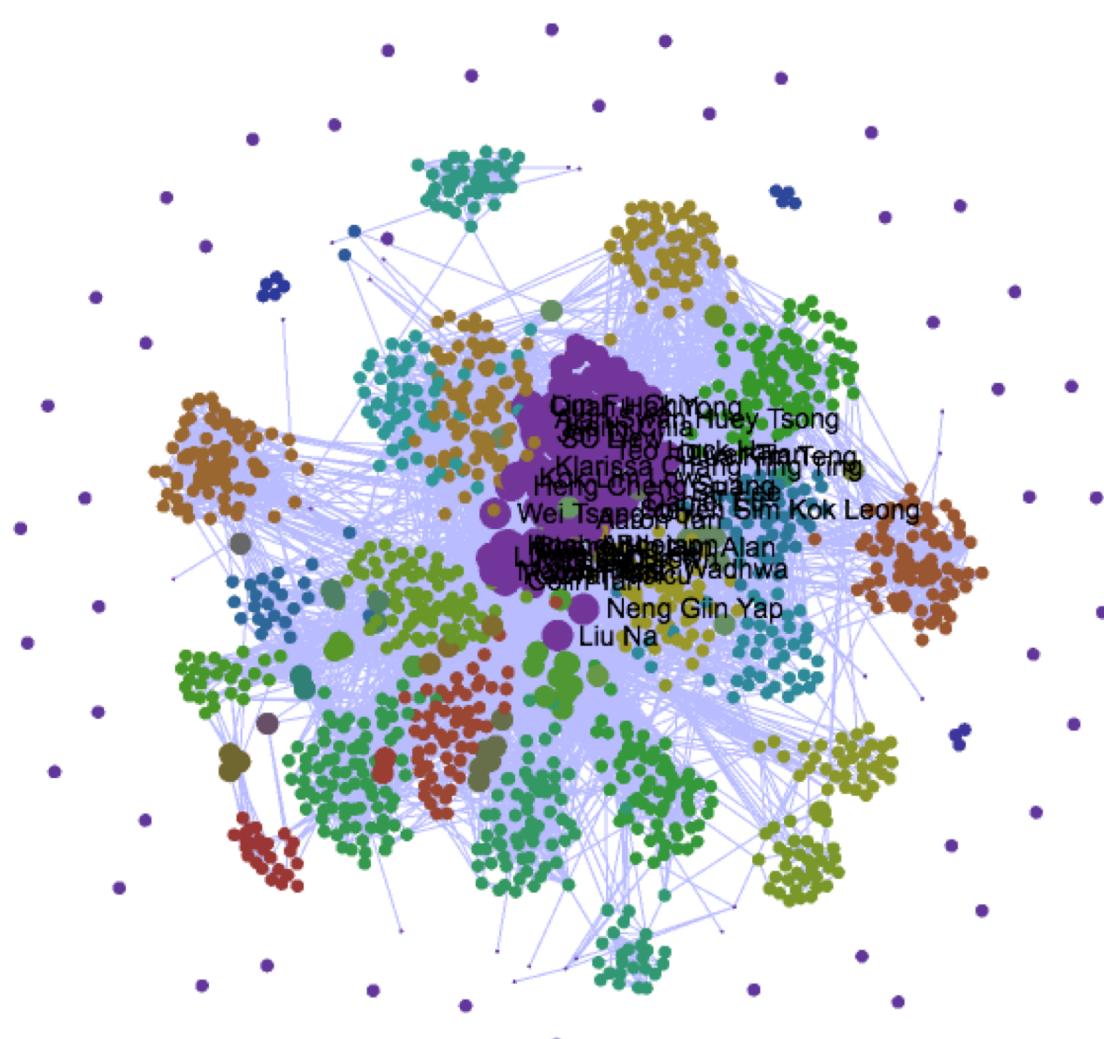
<http://pals.yjyao.com>

- See if you can discover some of  
“What you don’t know  
that you don’t know.

**Send me email at leonghw@gmail.com  
to give me feedback**

# Community of YuJian's friends





Elaine Chan

Aihui Ong

Rhandisee  
Singh

Wee Sun Lee

Stefanie  
Li-Nah Ng

Rina Selamat

William Ku

Michael Lin

Aaron Tan

Min-Yen Kan

Weehyong  
TokEng Pin  
Kwang

Sze Eng Koon

Lim Yan Hong

Tan Tiong  
Ghee

Raymond Ng

Communities of LeongHW's friends (1500 nodes)

# About YuJian's project...

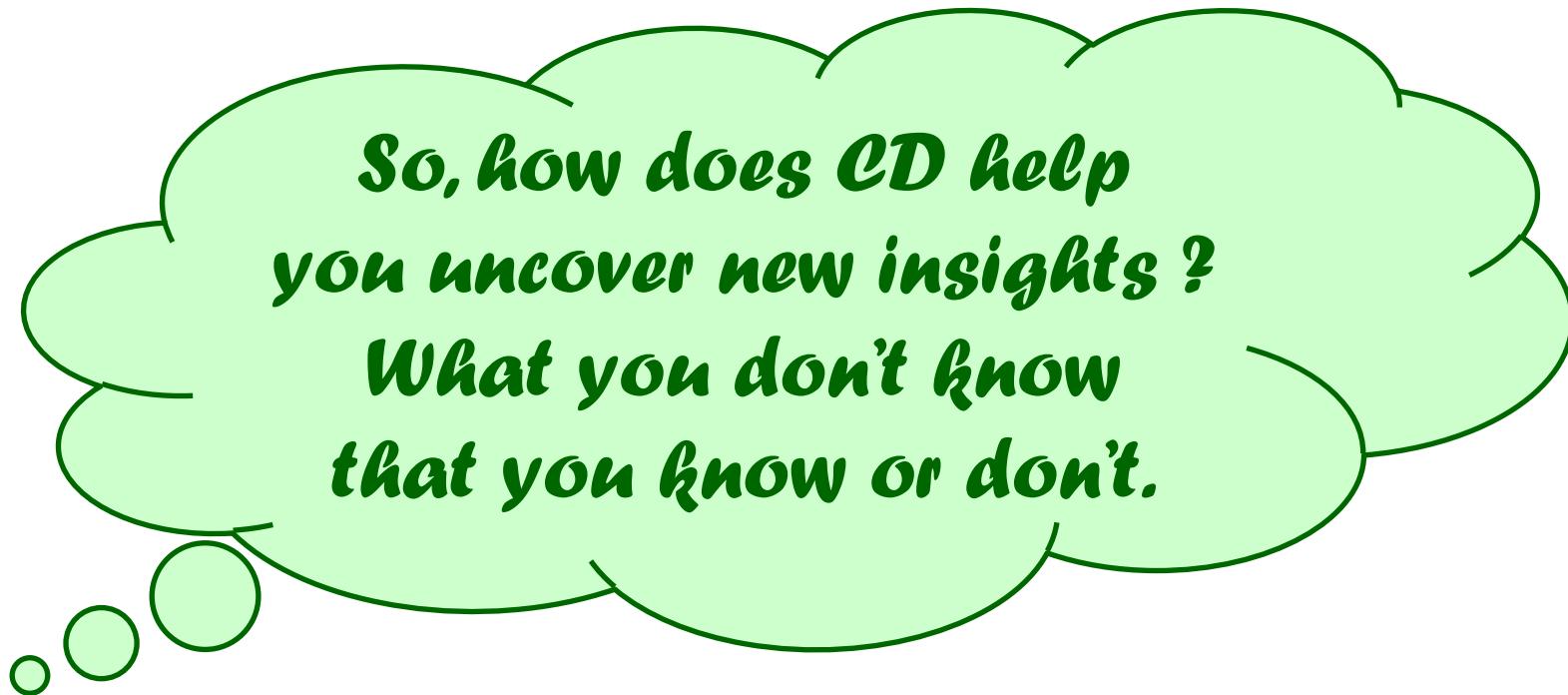
## Fun and Fulfilling Project

- ❖ Learns research on algorithms
- ❖ Develop a cool apps
- ❖ Work still in progress...

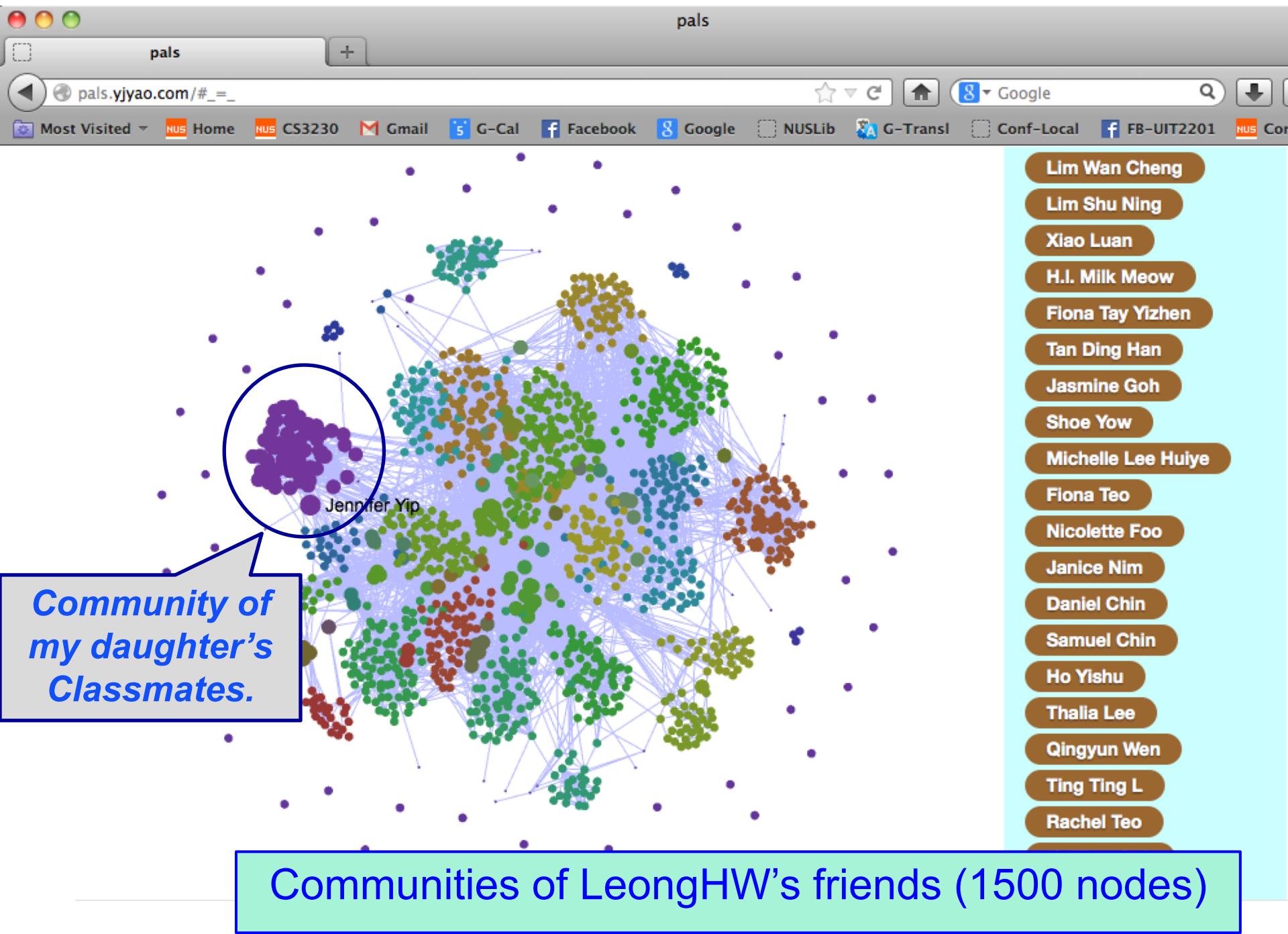
*That summer (2014),  
Yujian got internship in Dropbox*

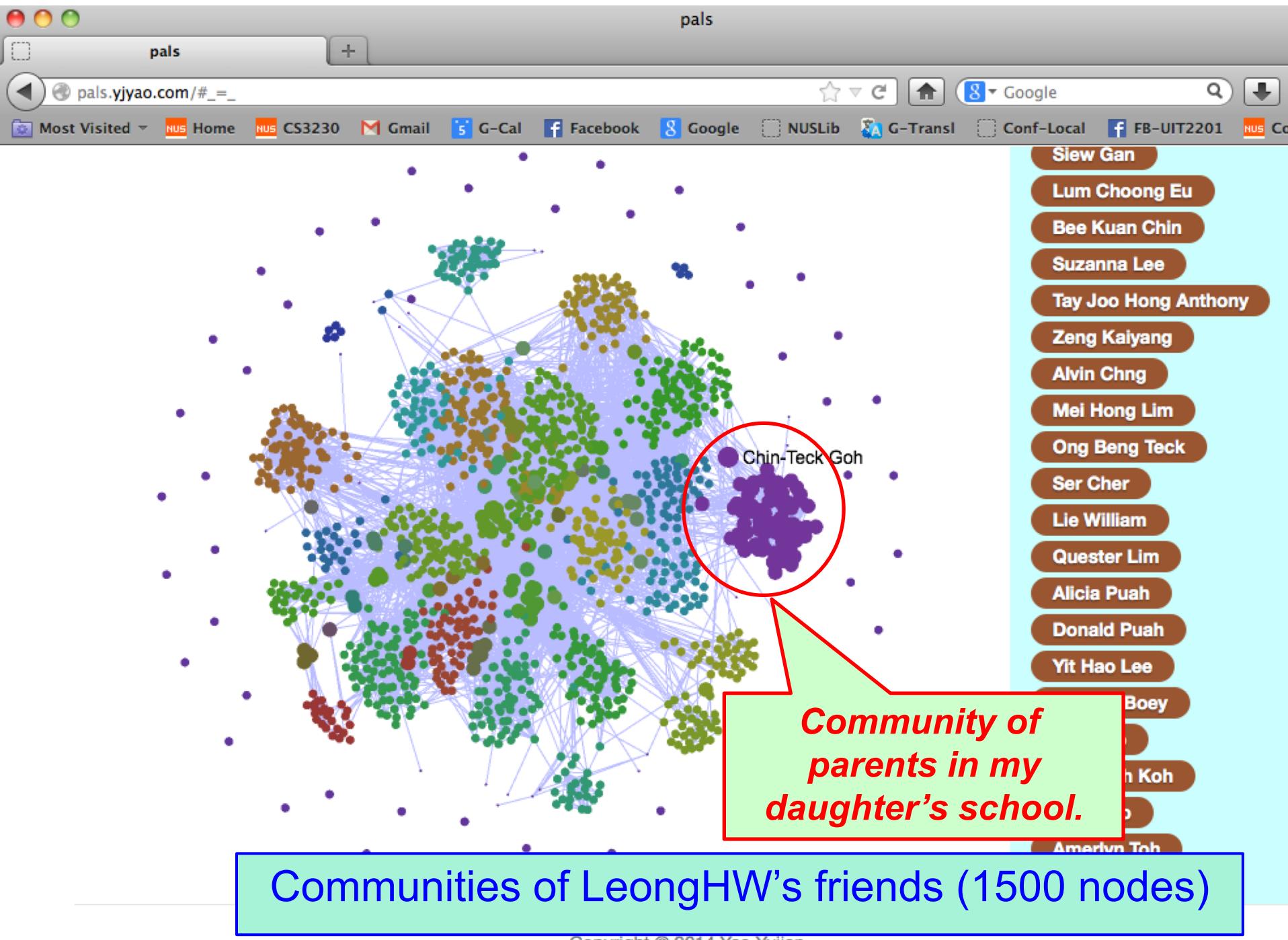
*implemented a new feature  
(can see it if you use the paid version)*

# CD helps develop new insights



I ran the system on my FB



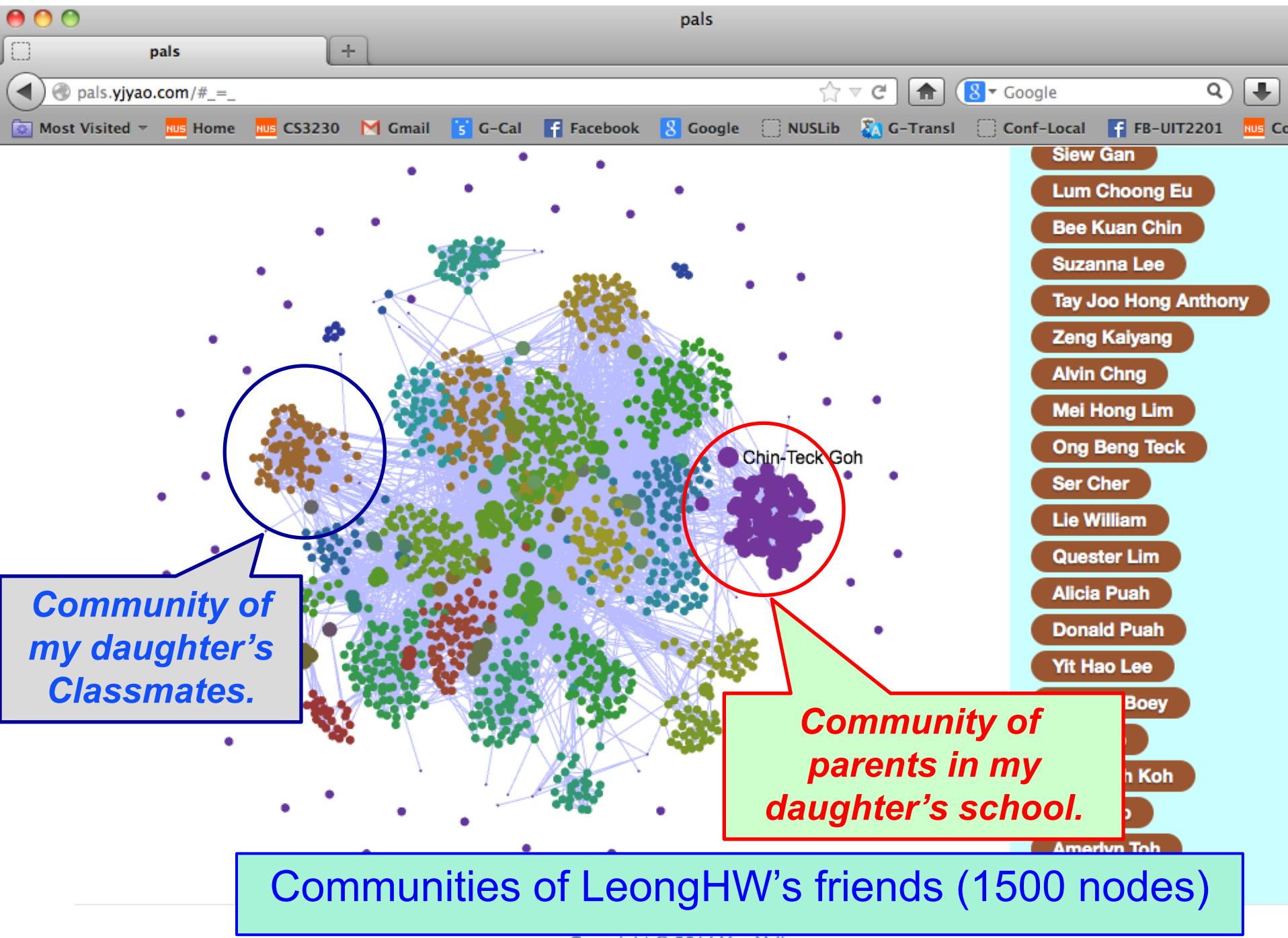


# CD confirming known insights

---

- **Children's Community** consists of classmates of my daughter
- **Parent's Community** consists of parents of the students

CD has confirmed insights that are  
“what I know that I know”



# New insights from CD

- Children's Community is far apart from the Parent's Community

**New insight:** The children are not FB friends of their parents.

This new insights makes a lot of sense.

CD gave new insight that is “what I don't know that I *might* know”

Most Visited

NUS Home

NUS CS3230

Gmail

G-Cal

Facebook

Google

NUSLib

G-Transl

Conf-Local

FB-UIT2201

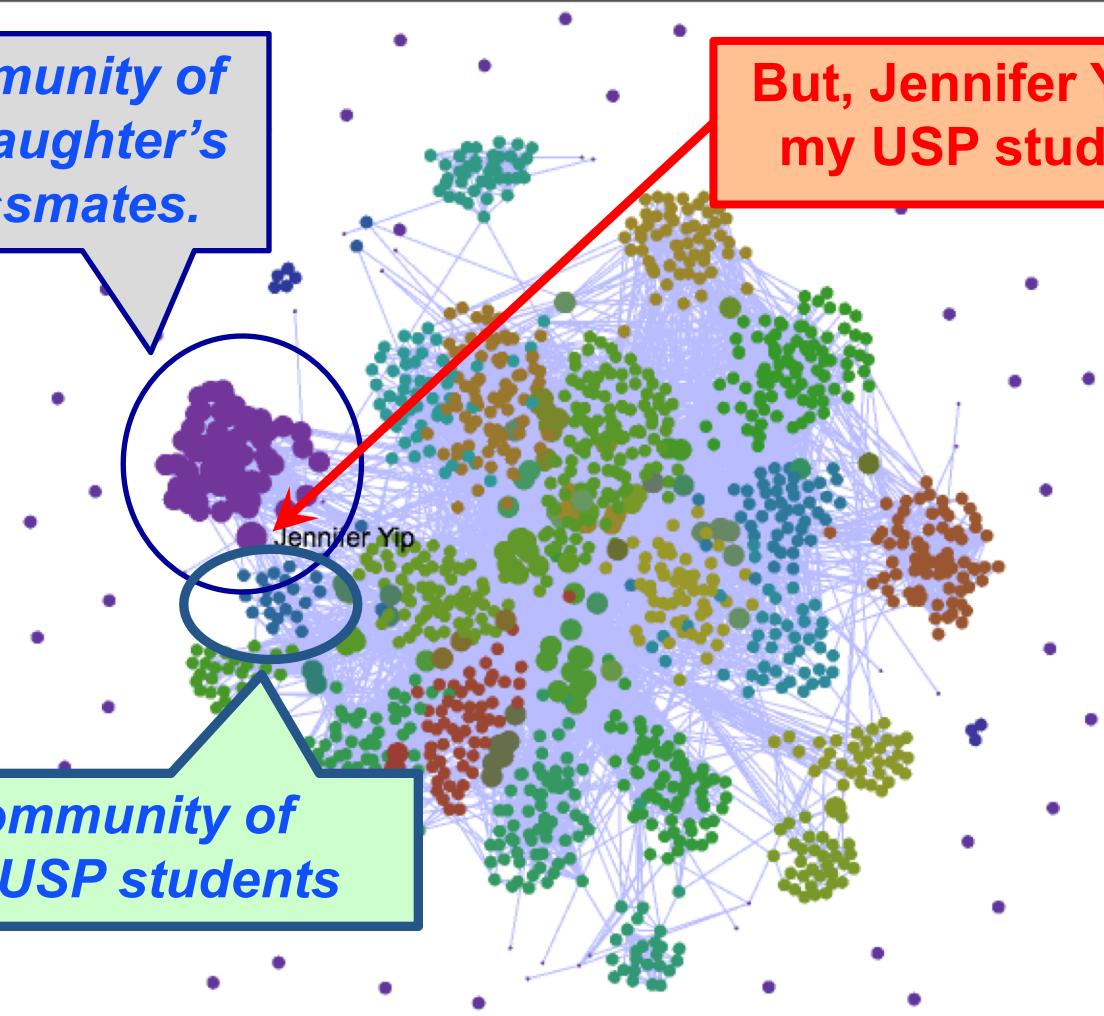
NUS

*Community of  
my daughter's  
Classmates.*

*Community of  
my USP students*

**But, Jennifer Yip is  
my USP students**

Jennifer Yip



Lim Wan Cheng

Lim Shu Ning

Xiao Luan

H.I. Milk Meow

Fiona Tay Yizhen

Tan Ding Han

Jasmine Goh

Shoe Yow

Michelle Lee Huiye

Fiona Teo

Nicolette Foo

Janice Nim

Daniel Chin

Samuel Chin

Ho Yishu

Thalia Lee

Qingyun Wen

Ting Ting L

Rachel Teo

HuiYan Ong

Communities of LeongHW's friends (1500 nodes)

# CD confirming known insights

- ❑ Jennifer Yip is my USP student. She is also 2 years senior of the Children community
- ❑ But Jennifer Yip is clustered with the Children community.
- ❑ WHY?

Need external confirmation data.

CD gave new insight that is  
“what I don’t know that I don’t know”

# Message Jennifer



Jennifer Yip

You're friends on Facebook

Studied at National University of Singapore

Lives in Singapore

05/09/2014 12:31AM

... 😊

Jennifer: Saw that you are FB friends with a few of my younger daughter's current/former classmates – Cheryl, Etsuko, Yan Li, Jamie, An Lin, etc. You still doing something with NYGH/HCI?  
Just curious. (I know them too.)  
--hon-wai

hi prof!

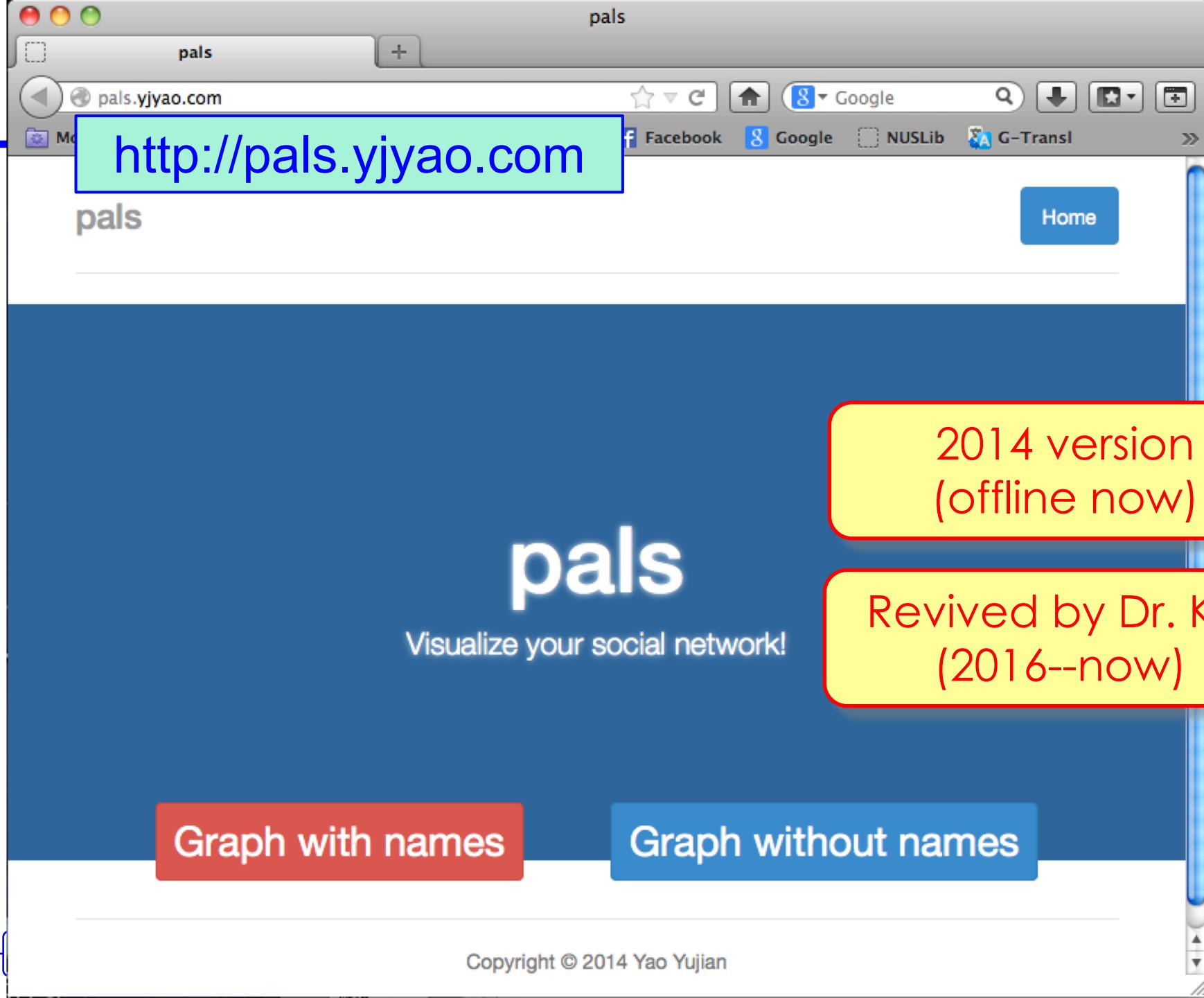
oh my, it's a small world

yes i trained them for a competition called the world scholar's cup



Hon Wai

is your daughter's name amanda leong...?



# But Pals is no longer functional?

---

“Hi Prof Leong,

Yes it's not functional because Facebook updated their API to disallow getting mutual friends out of any friend. Without that we can't construct a friendship graph.

According to this:

<http://techcrunch.com/2015/04/28/facebook-api-shut-down/#.yt0bn:nael>

it happened some time in April this year (2015).”

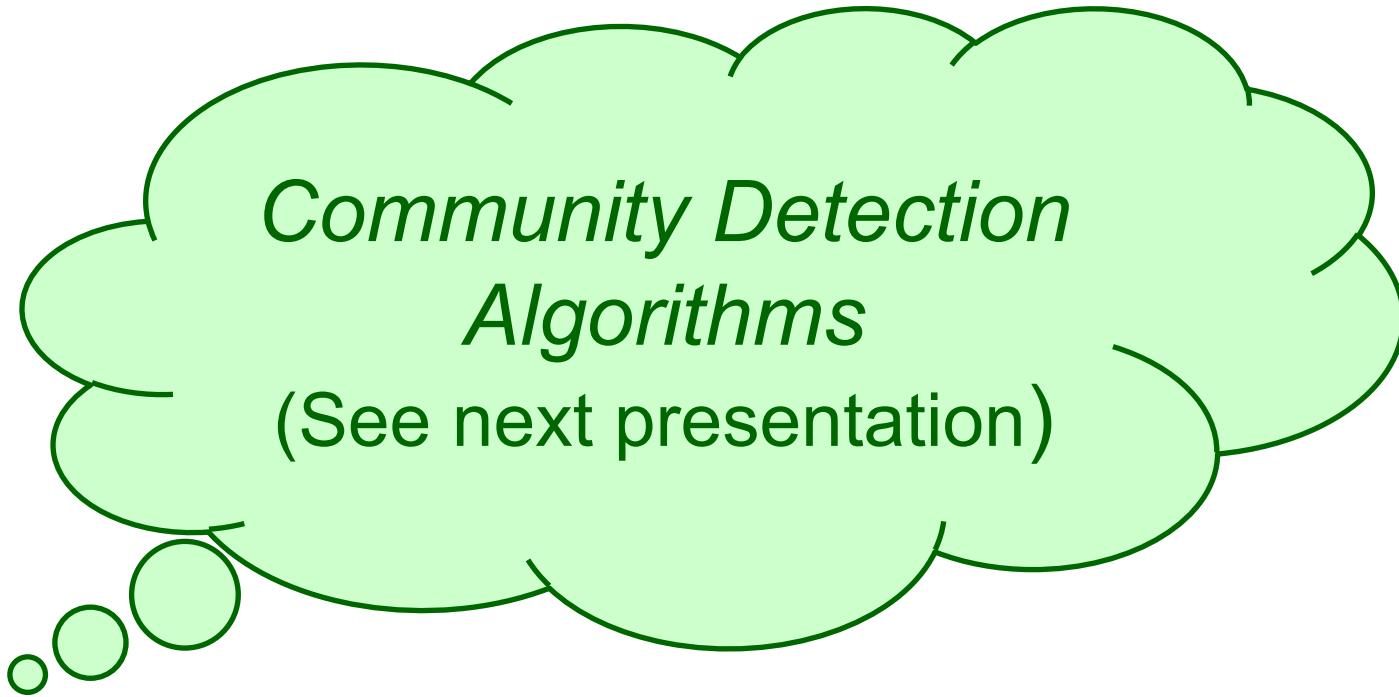
Hope this helps!

Cheers, Yujian

# Re-birth of Pals

---

- ❑ Need new picture,
- ❑ New url to link to...



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

# 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/>

*Thank you.*

## Q & A

Contact:  
**Hon Wai Leong** (梁汉槐)

FB, email: [leonghw@comp.nus.edu.sg](mailto:leonghw@comp.nus.edu.sg)  
<http://www.comp.nus.edu.sg/~leonghw/>