Cheating at Keyword Bingo

D.Vrodoljak and G.Kleemann Phd.

About us (BDSG)

BDSG = team of UCB DS instructors, and Alumni originally founded with the goal of bringing MIDS work to market, and later developing into:

- consulting/DSaaS biz

- working with a range of customers, including Fortune 500's

Project Pigeon

Have you ever tried to get a job interview with a resume?

Resume tuning

- What skills are employers actually looking for ?
- Which keywords are required for which jobs?
- Which keywords go together (complementary)
- Which keywords can be trimmed (interchangeable)

BD
SG

History

- TF-IDF, too simplistic
  - Hyperdimensional TF-IDF
  - A lot of operations
  - Actually a graph problem
- This is actually a graph problem
  - Moved to graph analysis using GRAKN.ai

BD
SG

Exchangeable relationships

- Which keywords can be trimmed (exchangeable)?
- Which of a pair matches with the target jobs?
  - R and Python C++ and JAVA

Complementary relationships

- Which keywords go together (complementary)
- Web development + JAVA

Data from actual job postings

- Keyword bank
- Keywords commonly associated
- Job title data and patterns

BD
SG

Graph and NLP analysis

- Multiple public data stores
- Job description ontologies
- Want ads
- Resume
- Collaboration records

Job related Data:

51,700
Job Skills on LinkedIn

—

20,000 (Jobs /major metro area /job title)

—

2,000,000
Relationships to score, per search, after cleaning data

Publication Data:

24,000,000
Full PubMed Dataset (all lit)

—

2,900,000 (~1.5 Million Open Source)
PubMed Central (BioTech -open access subset)

—

1,219,850
Total Entries, Condensed, after Cleaning/Pre-Processing

BD
SG

A mobile interface – for job seekers

- What are your resume keywords?
- What are you applying for?
- What other titles should you consider?
- Terms that should be added?
- Terms that should be removed

BD
SG

A mobile interface – for recruiters

- What are the current trends in hiring?
- How does a candidate match up?
- What should they get better at?
- Are there other types of jobs to consider?

BD
SG

A mobile interface – for employers

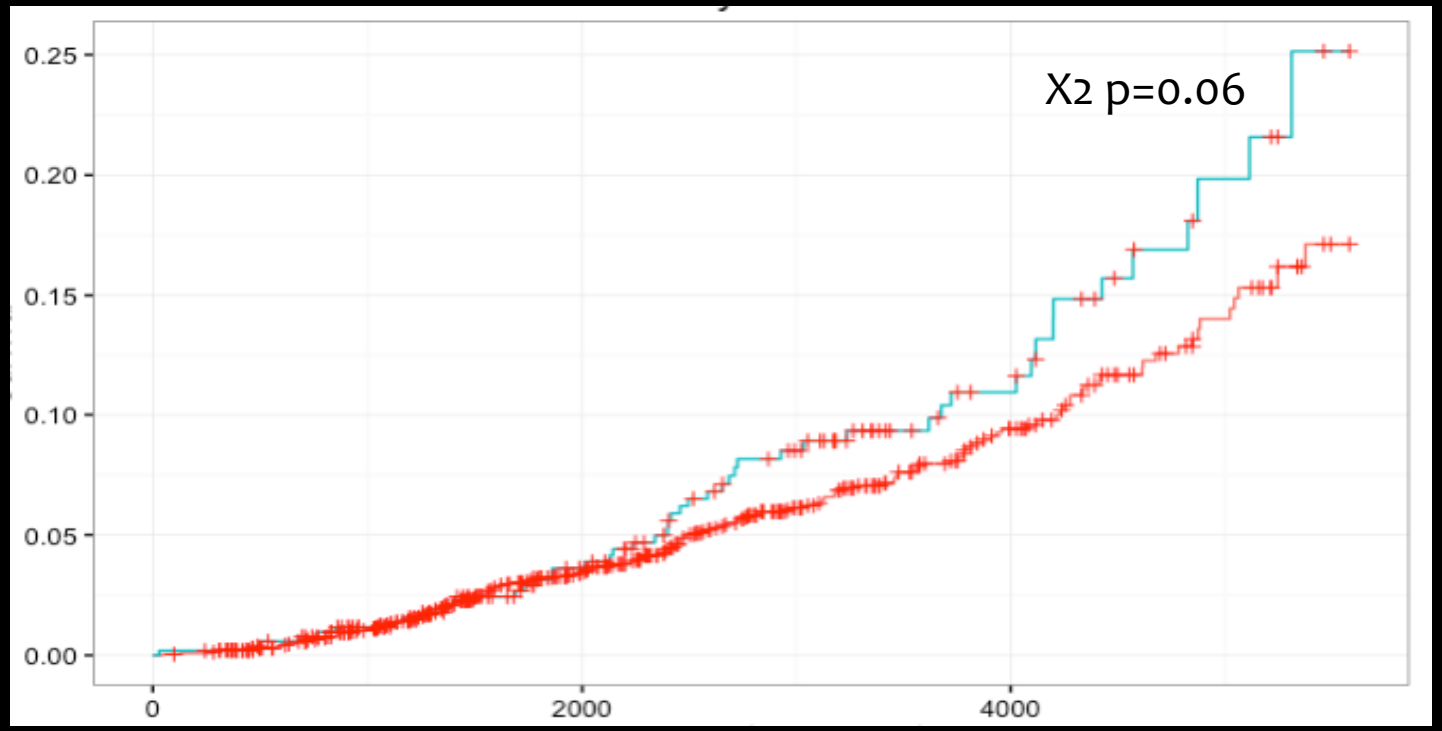
- Which domain is the scientist candidate really an expert in?
- Are they a good collaborator?
- What do their collaborators know?

BD
SG

Scilit

Have you ever wondered what type partner a new collaborator will be?

Teams with more scientists IPO faster



> 50% scientists

fewer or no scientists

BDSG biotech sector analysis BioPredictor

BD SG

But how do we judge scientists?

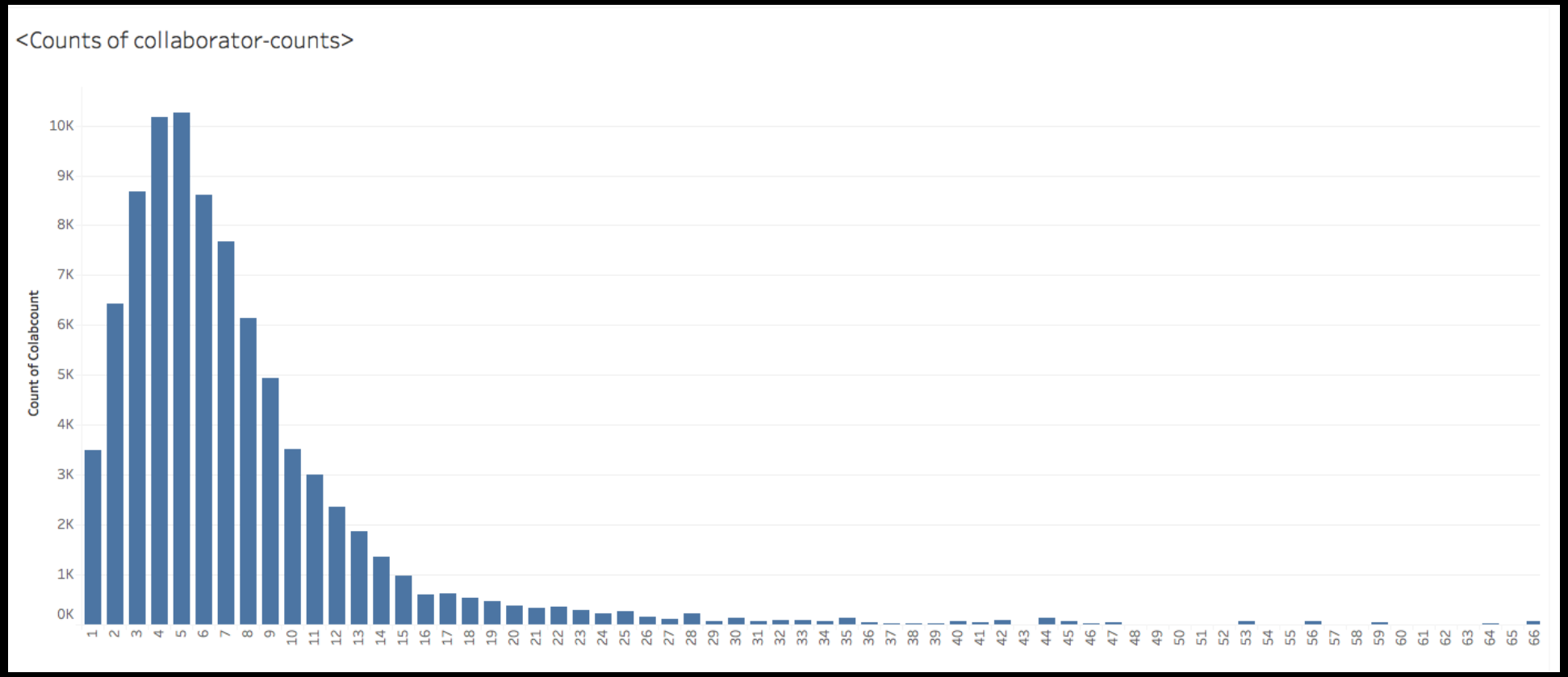
- Paper count
- Journal impact factor
- Citations
- Inside information
- Interviews

** Either expensive or basic **

What about their true expertise?

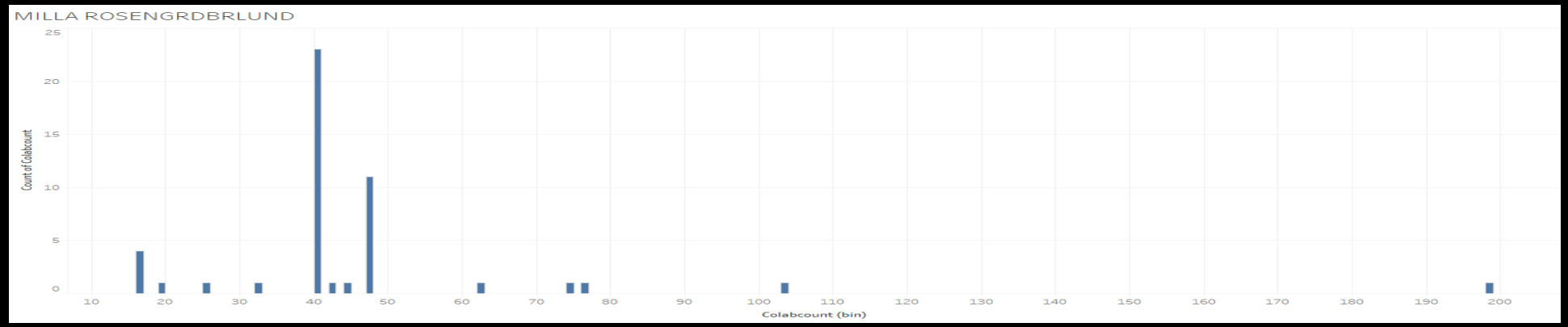
What about their network?

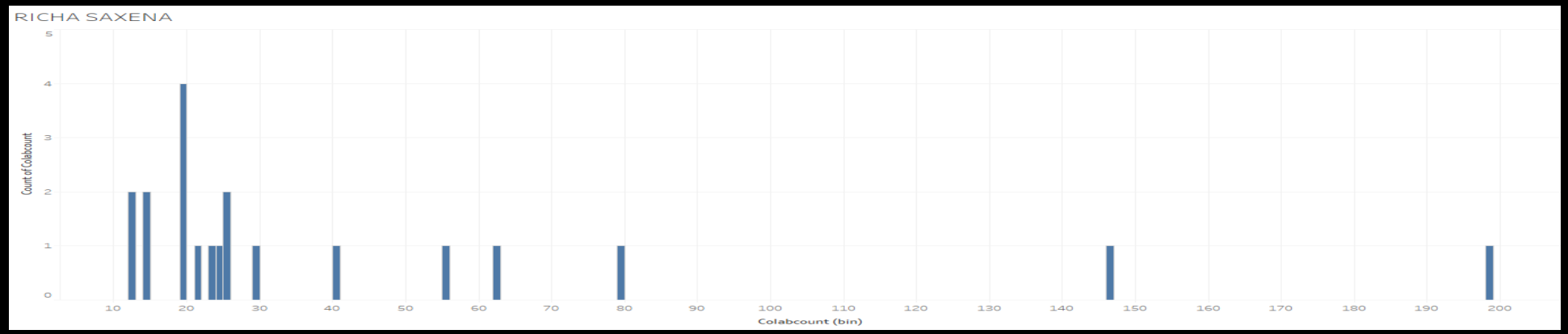
Paper counts per scientist globally



Individual's collaborator density distribution

of collaborators



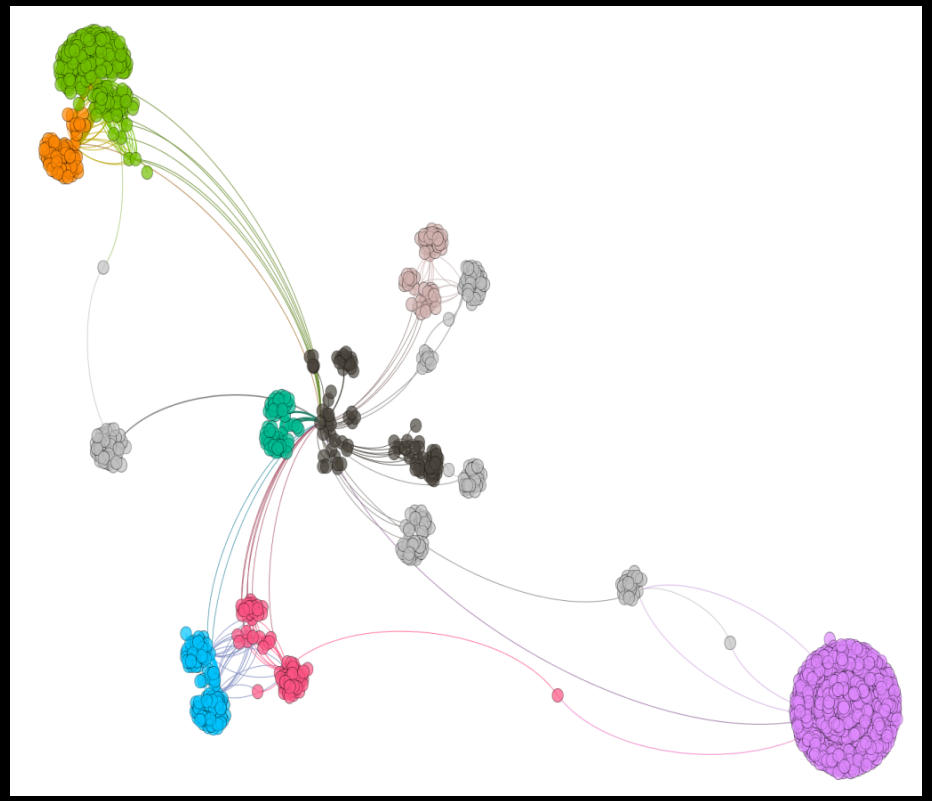


paper counts per collaborator

BD SG

Analysis of publication record can give answers

- A rich record of behavior
- Collaboration record
- Rich long term record of performance
- Record of alliances
- Expertise is coded in the network



Building the collaboration graph in 3 layers

- Person 1

Primary

Secondary

Tertiary

- 

Primary paper

Secondary paper

Tertiary paper

https://www.iconfinder.com

BD SG

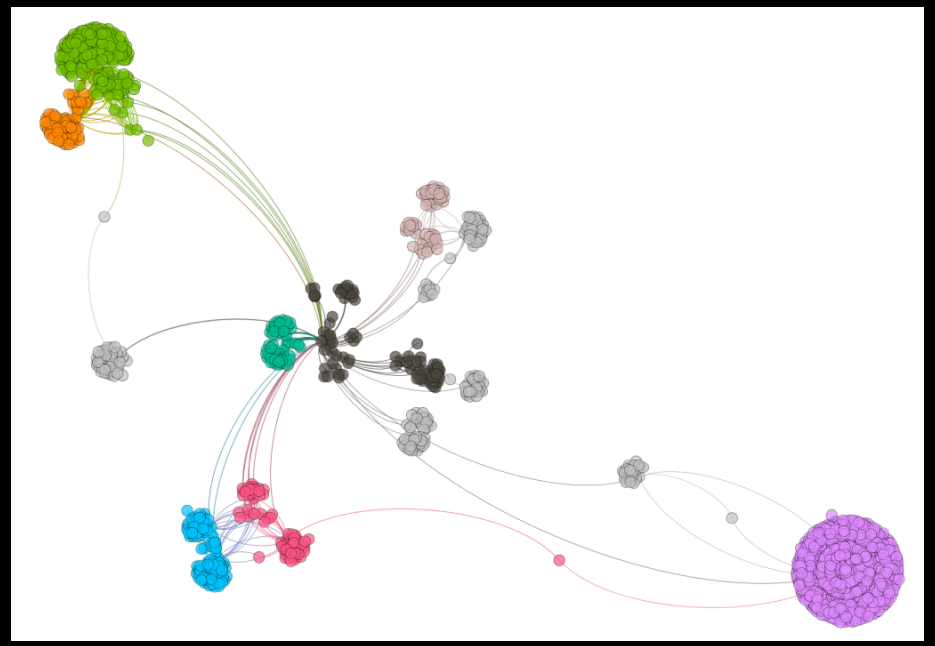
Consider the most prolific publishers

n	COUNT(r)
Node[487153]{name:"THE "}	2671
Node[865318]{name:"GUNTHER EYSENBACH"}	2055
Node[78005]{name:"HOONGKUN FUN"}	1329
Node[1520956]{name:"SEIK WENG NG"}	1233
Node[12886]{name:"WEI WANG"}	1203
Node[33084]{name:"WEI ZHANG"}	1029
Node[148]{name:"YAN LI"}	829
Node[6653]{name:"WEI LI"}	778
Node[15555]{name:"JING WANG"}	737
Node[20702]{name:"LI ZHANG"}	670

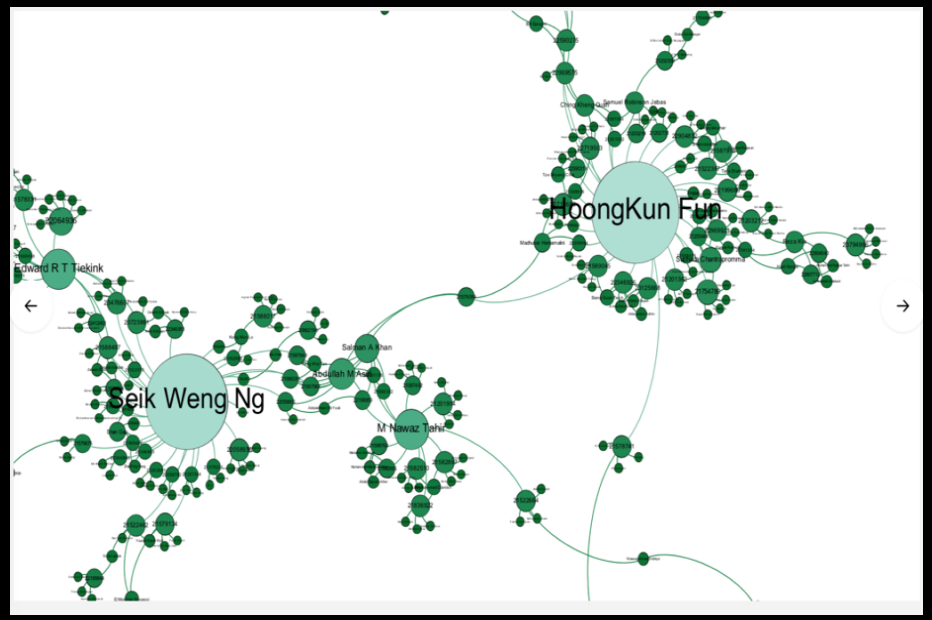
10 rows
24526 ms

Different collaboration patterns

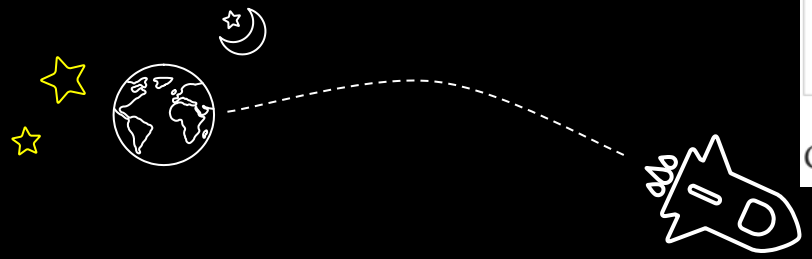
Gunther Eysenbach



Seik Weng Ng and HongKun Fun

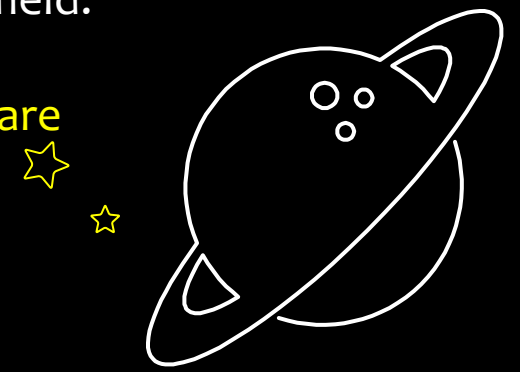


A Big publisher with 2055 links!



“one of the most productive researchers, editors, and publishers in the online health field.”

in 2004 received the Janssen-Cilag Future Award, referred to as the German “health care nobel prize”.



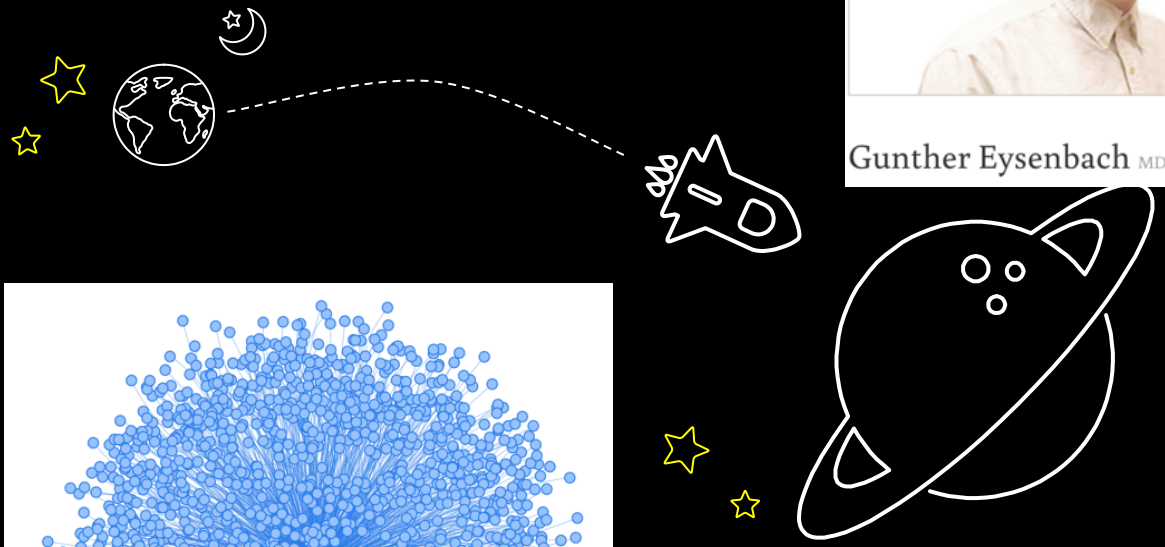
Founder of an academic field!

- association between search engine queries and influenza incidence,
- He coined the terms **"infoveillance" and "infodemiology"** for these kinds of approaches.

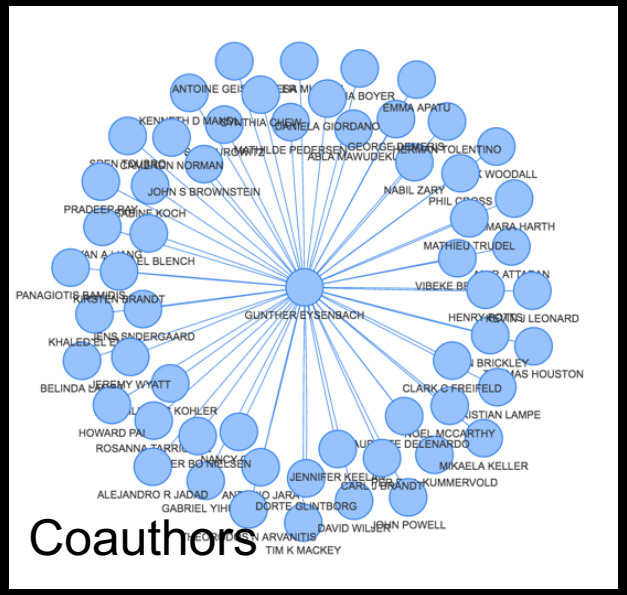
Source: http://ehealthinnovation.org/people/gunther-eysenbach/

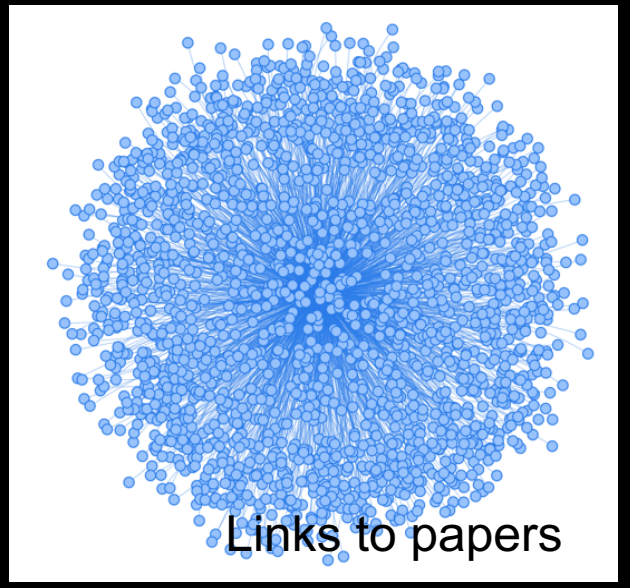
BD SG

Prof Eisenbach has a pattern that is interesting



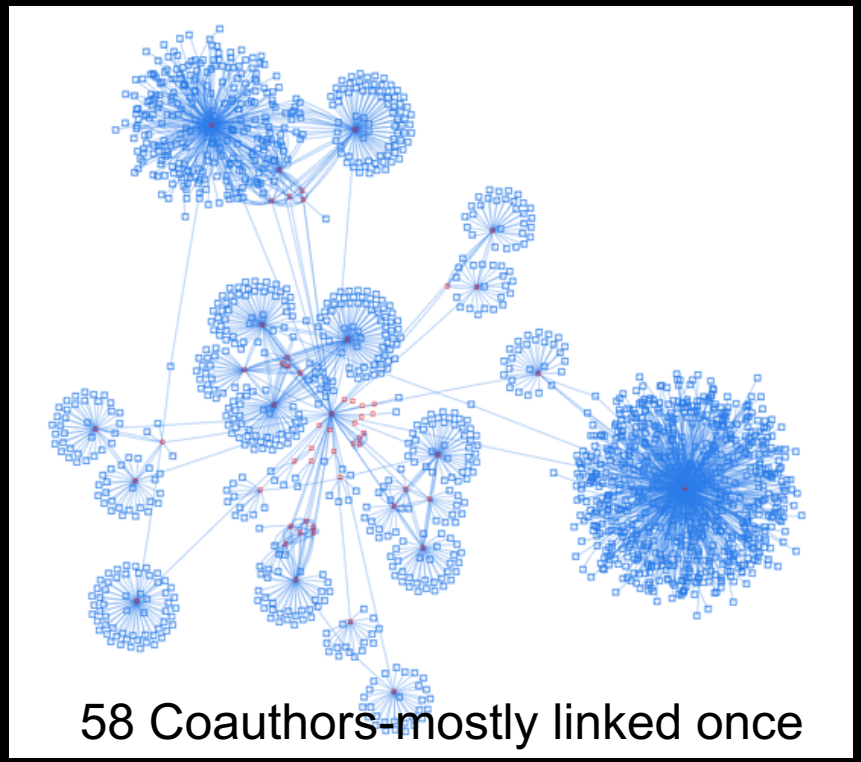
Many collaborators
Few publications

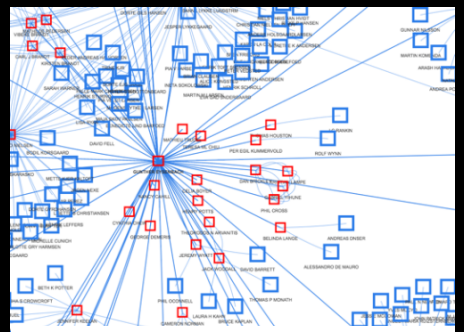




BDSG

Limited collaboration and a sparse network

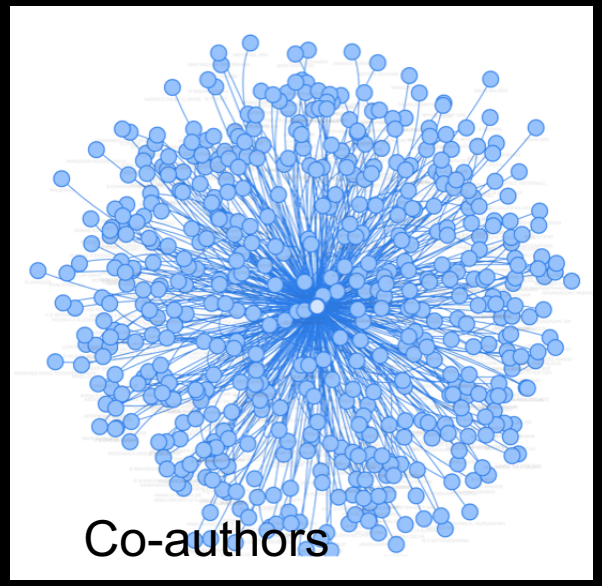

58 Coauthors-mostly linked once

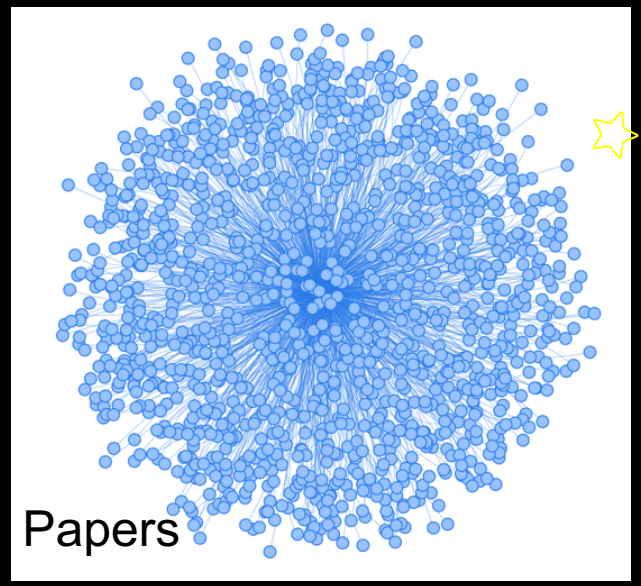

primary connections in red

About half of his papers appear to be written by him alone!

Prof Kun has a pattern closer to what we expected

Many collaborators and many publications

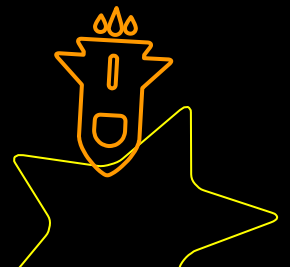




But there is there is something missing?

Gunther is a star, but he is only cited for 120 papers and 40 book chapters.

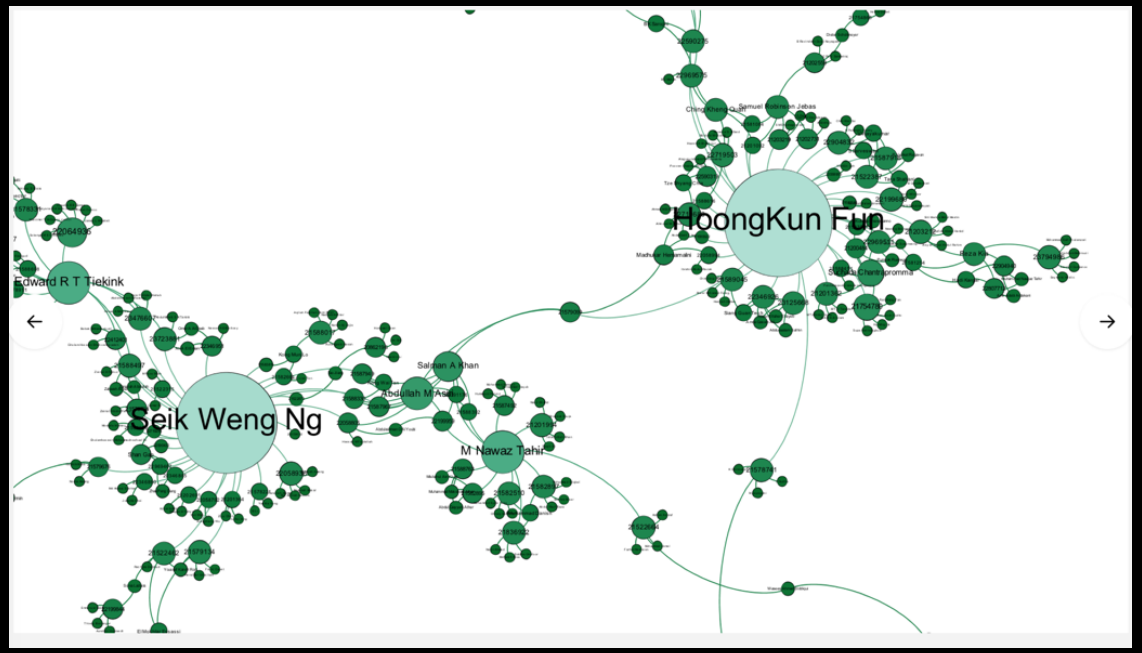
So what are those other 1900 links?
* Edited papers



BD SG

The NEW biggest publisher

- Highly collaborative
- Has co-authored > 2000 papers
- Independent identification from sub-setting and total dataset approach

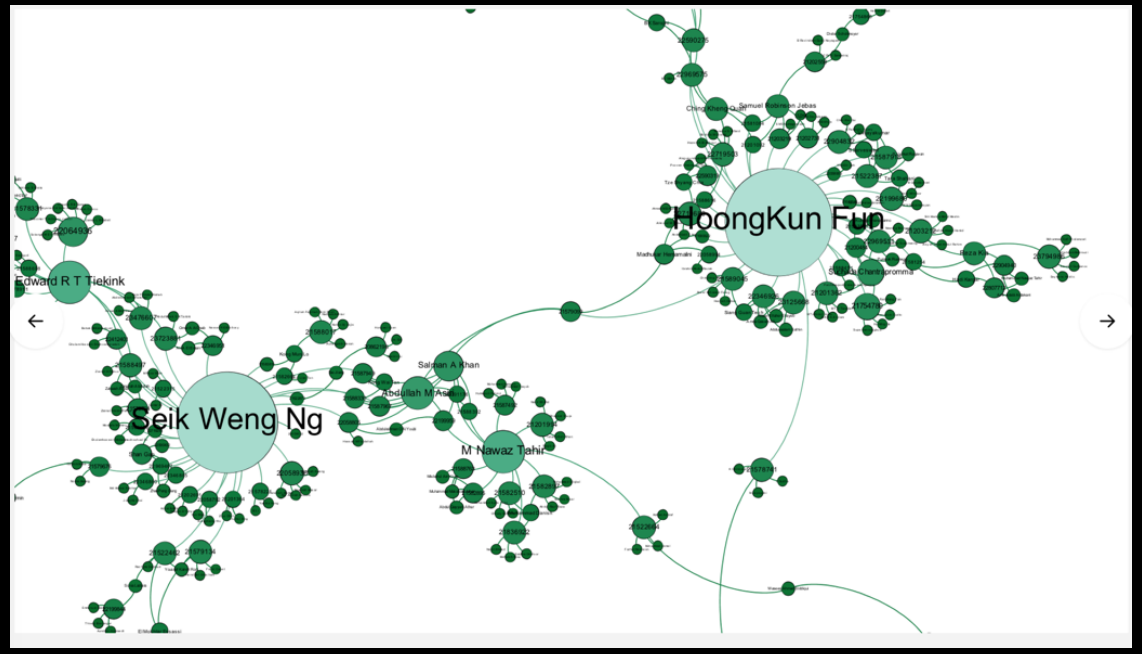


Highly collaborative X-ray crystallographer

- *Prof. Fun has published more than 2618 Science-Citation-Indexed (SCI) papers*
- *collaborating mainly with researchers from China,Taiwan, Thailand, India, Pakistan, Iran, Iraq, Japan, Egypt and Saudi Arabia. Under his supervision, this research group has solved a few thousand structures.*

There is more to the Prof. Kun Story

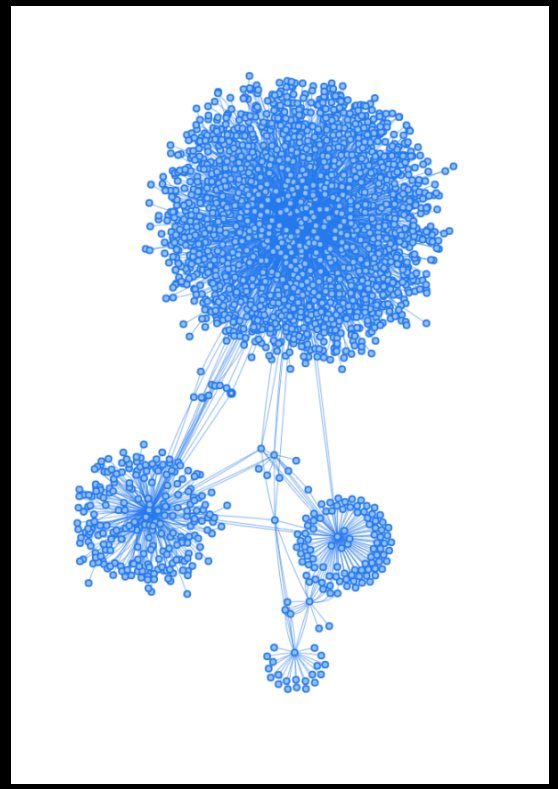
- **The two of the biggest publishers look very similar**



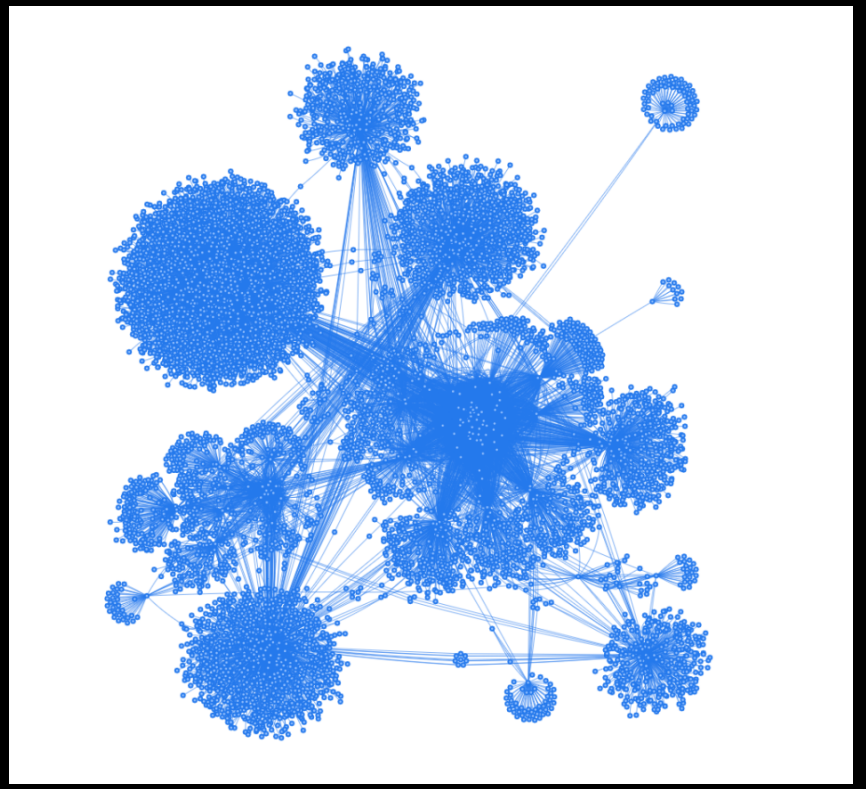

```
{name:"GUNTHER EYSENBACH"} | 2055
name:"HOONGKUN FUN"}       | 1329
]{name:"SEIK WENG NG"}     | 1233
```

Given some networks of interest

Scott W Emmons



Coleen Murphy

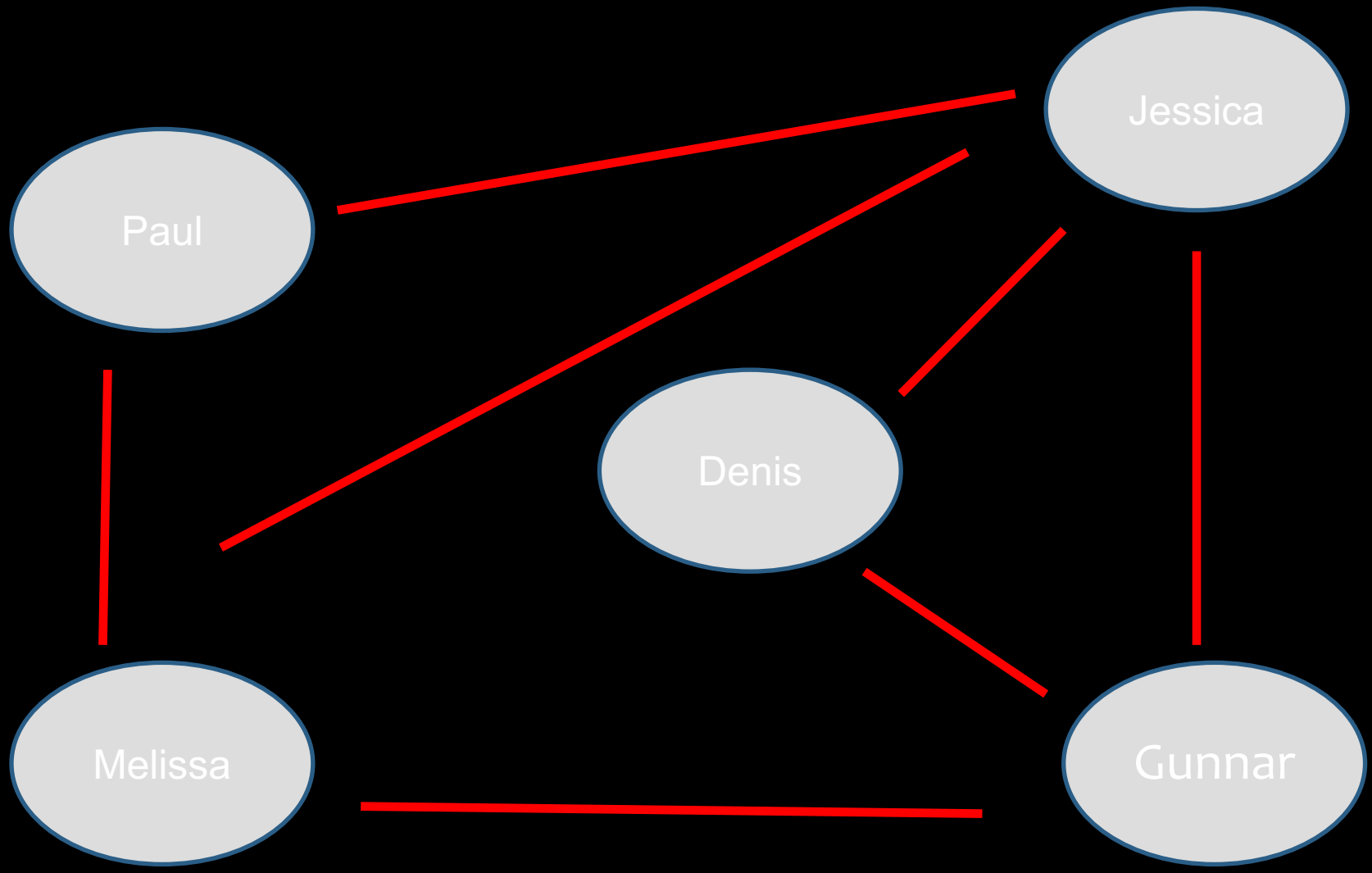


BD
SG

How can we assess these networks systematically?

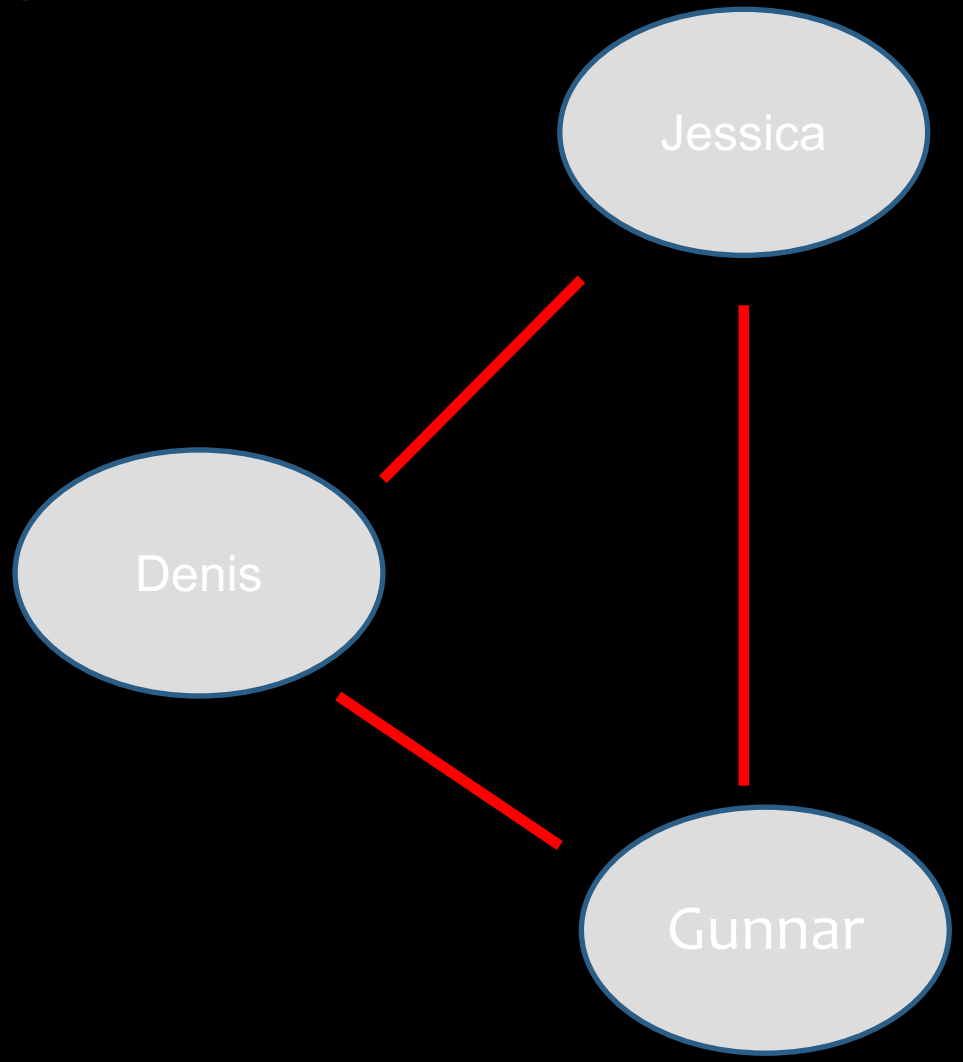
Definitions

Consider a simple social network

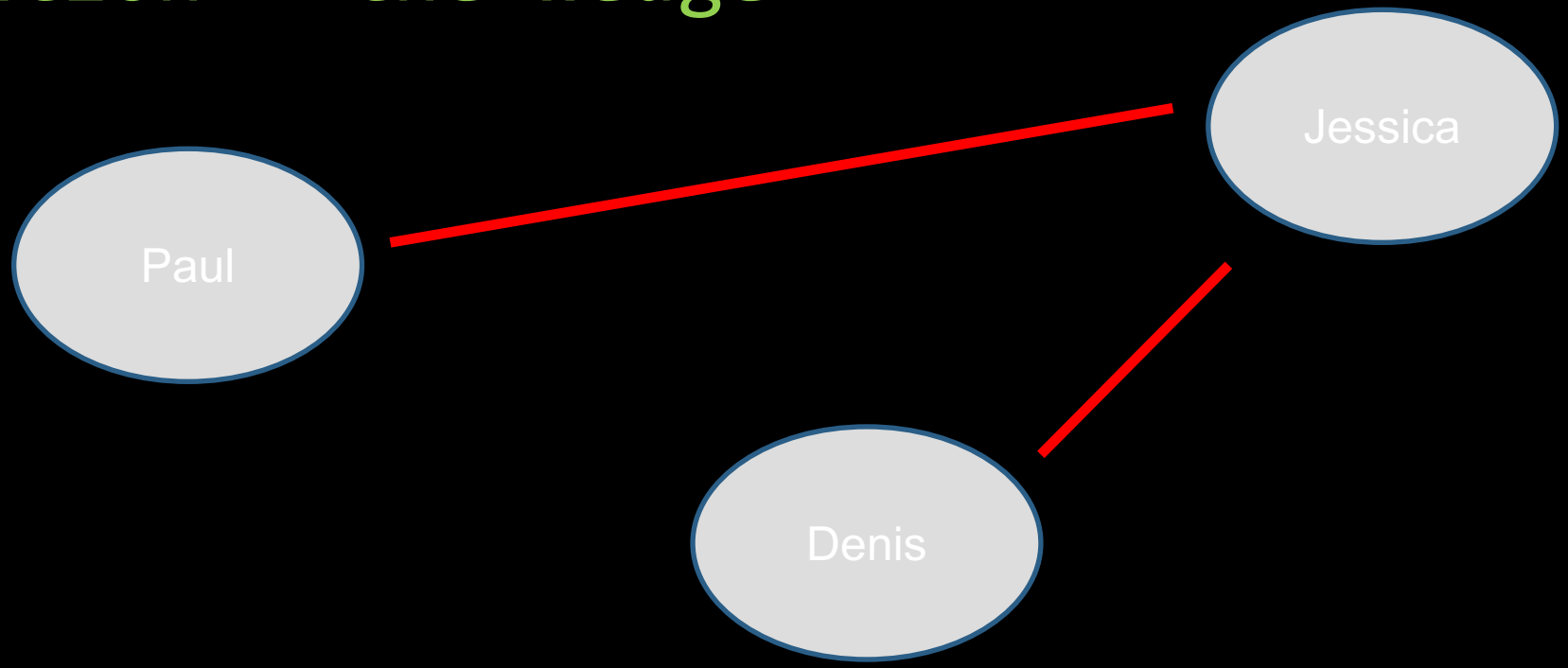


BD
SG

Definition – Triangle

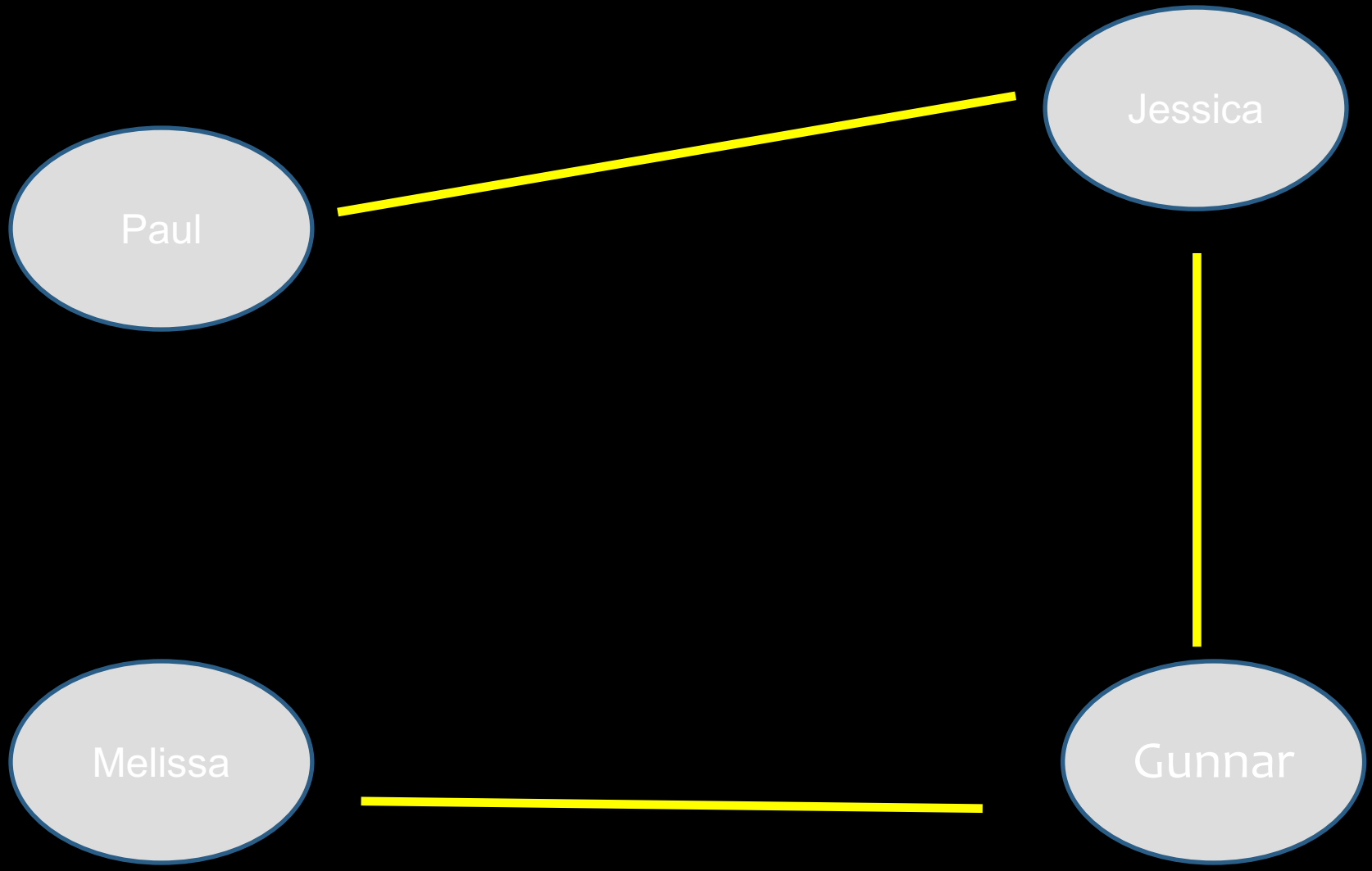


Definition – the wedge



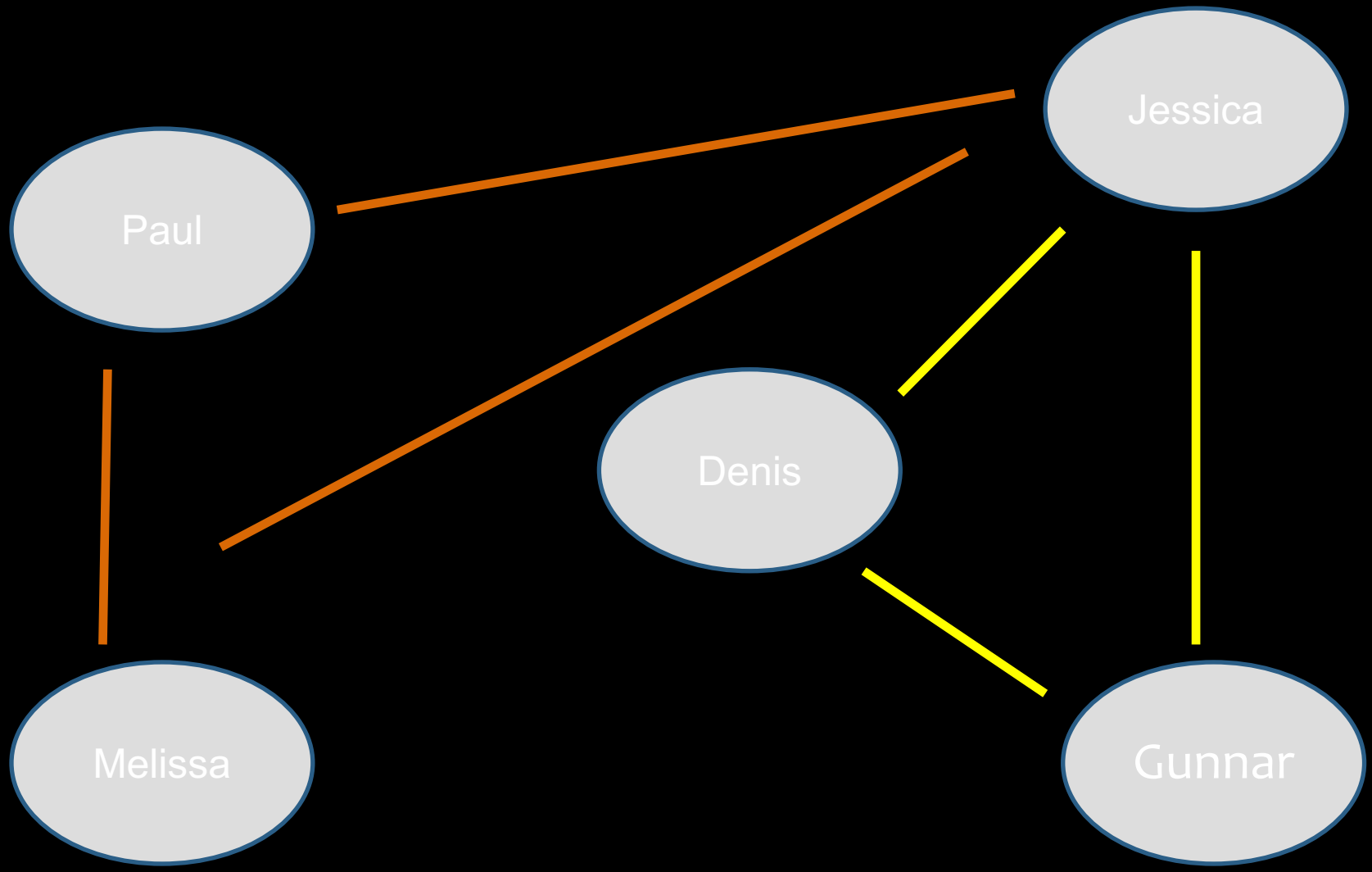
BD
SG

A wedge - third is not self



BD
SG

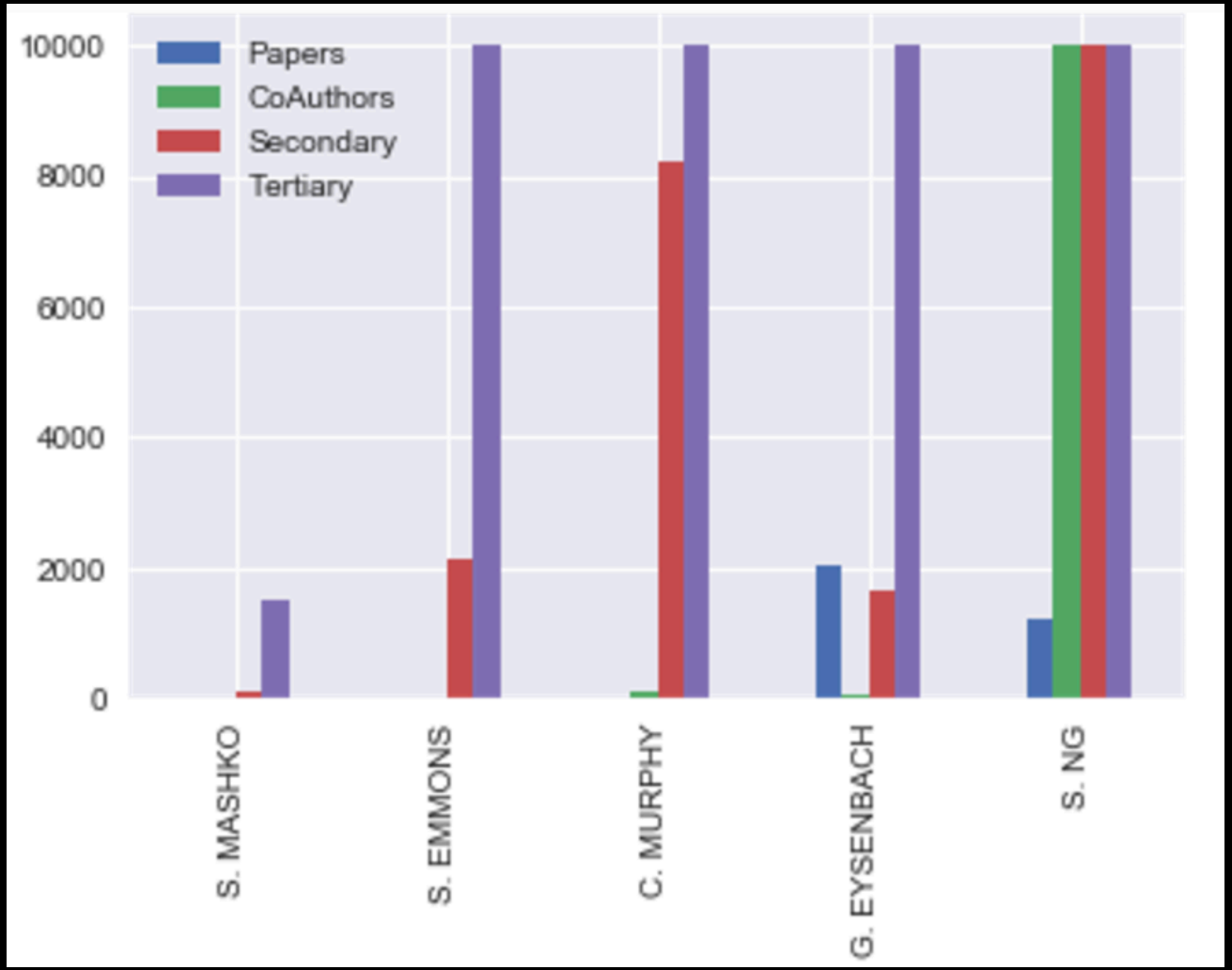
Two triangles – third is self



BD
SG

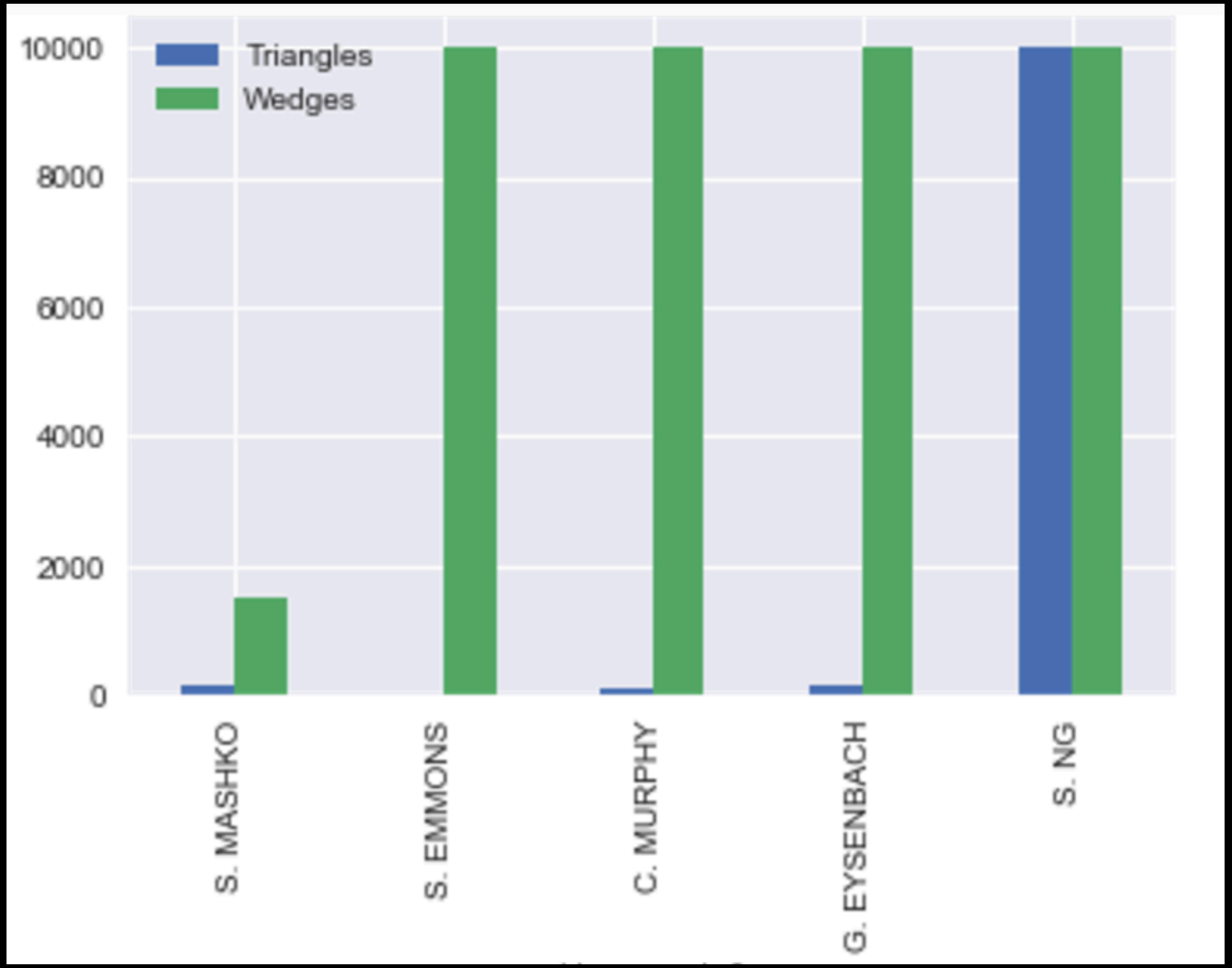
Graph traversal-collaboration patterns

Paper count at three levels



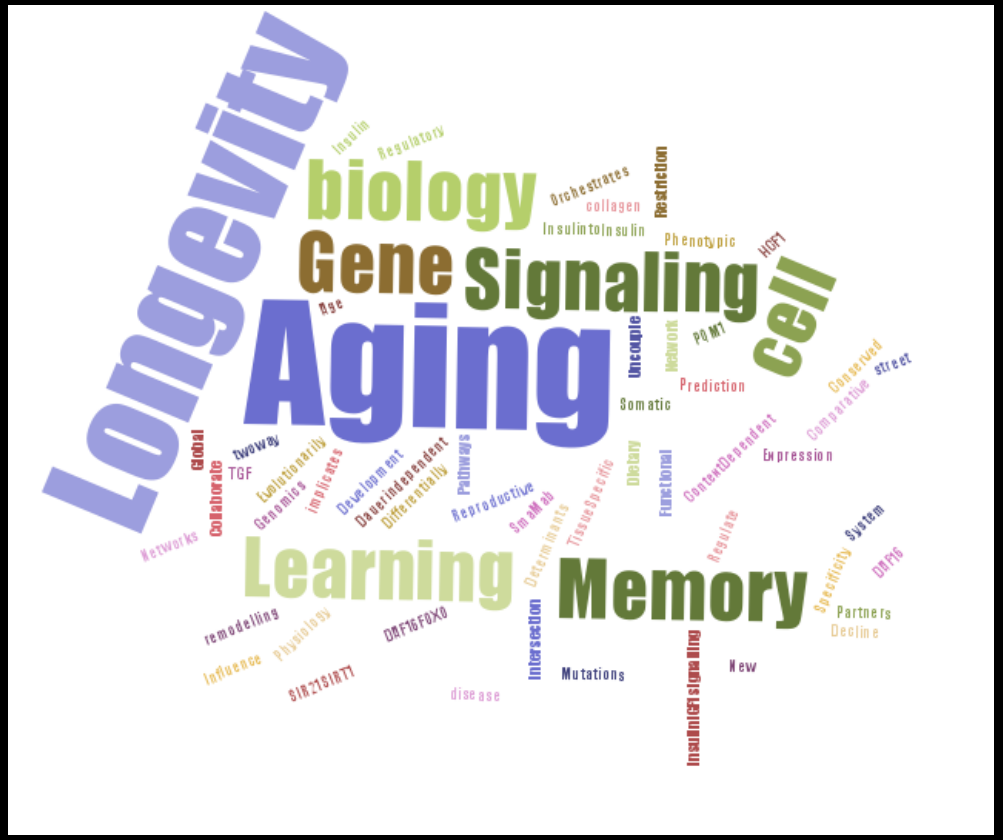
Graph traversal-collaboration patterns

Triangles vs Wedges



BD SG

Title clouds – what do they know?



- https://www.flickr.com/photos/utasel/6961732976

BD SG

Title clouds – what do their friends know

Secondary title cloud

Tertiary title cloud

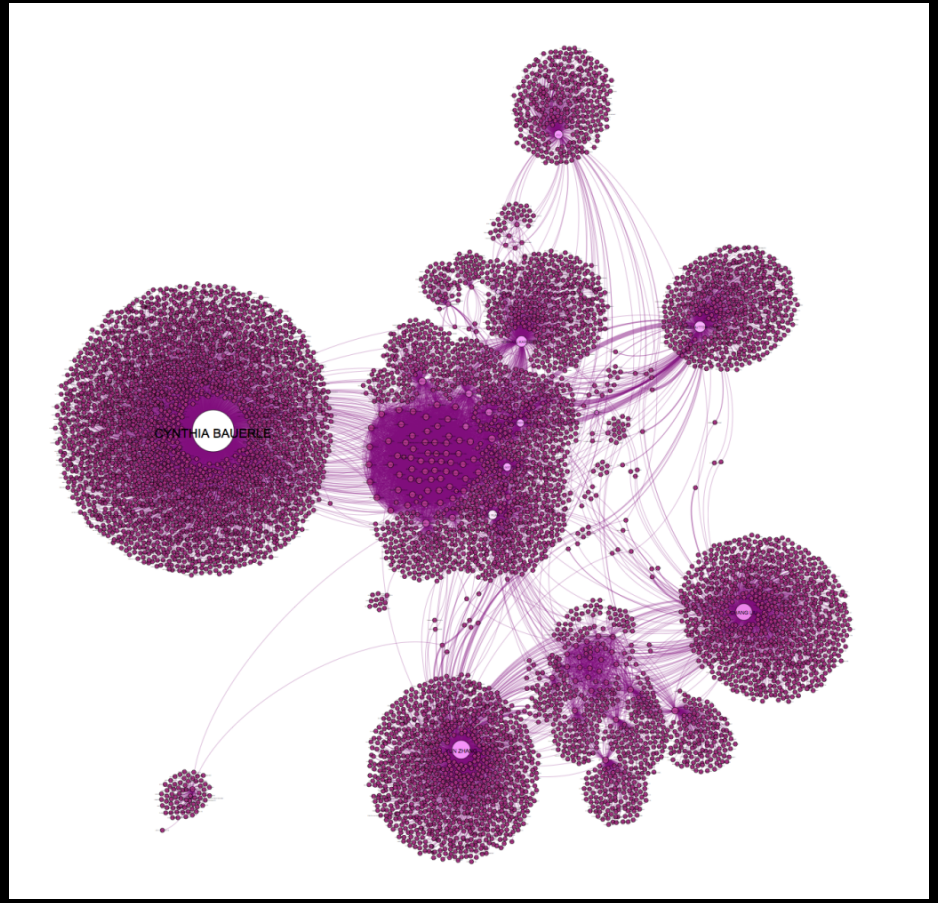
BDSG

Collaboration Statistics

Collaborators
Primary : 75
Secondary : 8000
Tertiary : <10000

Papers
Primary : 55
Secondary : 2000
Tertiary : 5000

Network
Wedges to triangle ratio: >10000
Title word count (primary): 235



Ongoing projects and collaborations

SciLit

MarketGraph

BioPredictor

Project “Pigeon”

ParlourBoard

Red Sands Robotics

About us (Speakers)

Denis Vrdoljak, MIDS
Managing Director, BDSG
Data Science Instructor, UCB
denis@bds.group

Gunnar Kleemann, PhD, MIDS
Senior Data Scientist, BDSG
Data Science Instructor, UCB
gunnar@bds.group

Thank you

Questions?