
YAHOO! KICKSTART
LAND OF DEFUNCT SOCIAL NETWORKS
BAHU
SOUNDPEDIA
YAHOO! MASH
YAHOO! 360
SEA OF FORGOTTEN MEMES
RECEDING GLACIERS OF AOL AND WINDOWS LIVE
AIM TUNDRA
BEBO
117 MILLION*
YAHOO
ORKUT
100 MILLION*
EBAY
GOOGLE INFORMATION GATHERING OUTPOST
FRIENDSTER
115 MILLION*
FARMVILLE VALLEY
DRUNK PICTURES MOUNTAIN RANGE
LAKE OF MYSPACE BANDS
FACEBOOK
500 MILLION
FORMER KINGDOM OF MYSPACE
125 MILLION*
NOT ALL USERS ARE ACTIVE
@FJORDS
OMFG WTF TTYL OCEAN
TWITTER
114 MILLION
DEATH VALLEY OF JOHN MAYER TWEETS
GOOGLE INFORMATION GATHERING OUTPOST
YOUTUBE
STRAIT OF RICK ROLL
ISLES OF REACTION VIDEOS
CAPE OF HITS REMARK
ISLAND OF TROLLS
SEA OF DESPERATION
*AS OF 2009
Hi5
70 MILLION*
STRAIT OF BING
HABBO
178 MILLION
SEA OF PERSONAL INFORMATION
MATCH
EHARMONY
CHEMISTRY
THE THIN LINE STRAIT
YOUPORN
ADULT FRIENDFINDER
LIVE JASMIN
FLIXSTER
BAY OF TAGS
FLICKR
VOLCANIC ISLAND OF IPHONE APPS
ISLES OF APPS
FOURSQUARE
DIGSBY
TWEETDECK
TWITPIC
STUMBLEUPON SEA
SEA OF MASTER DEBATERS
WORDPRESS
LIVEJOURNAL
BLOGSPOT
TUMBLR
LAKE OF CAKE WRECKS
BLOGSTER
LAST.FM
PANDORA
IMEEM
LOL CATS
DIGG
EMPIRE OF GOOGLE

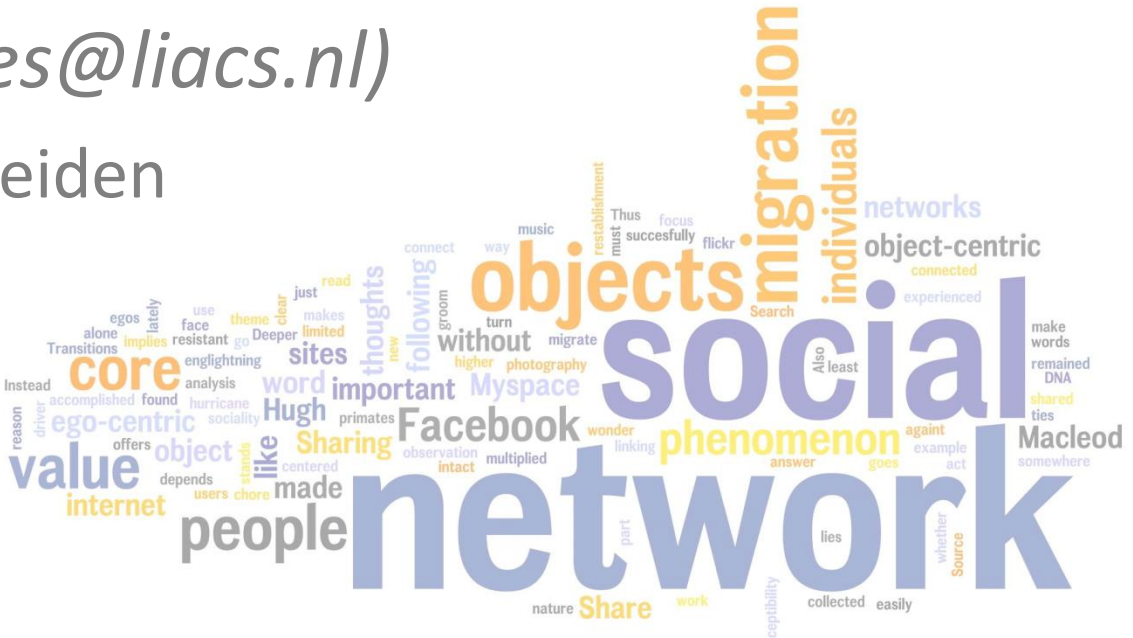
Universiteit Leiden

# Social Network Mining

Data Mining – November 11, 2013

*Frank Takes (ftakes@liacs.nl)*

LIACS, Universiteit Leiden



# Overview

Universiteit Leiden

- Social Network Analysis
- Graph Mining
- Online Social Networks
- Friendship Graph
- Semantics
- Example
- Conclusions

# Social Network Analysis

Universiteit Leiden

- **Social Network Analysis**: the study of social networks to understand their structure and behavior.
- **Social Network**: a social structure of people, related (directly or indirectly) to each other through a common relation or interest.
- Social Networks != Social Media

# Social Network Analysis

Universiteit Leiden

- **Social Network Analysis (SNA)**
  - Sociology
  - Algorithms
  - Data Mining
- **Social Networks**
  - Real-life (explicit)
  - Online (explicit)
  - Derived (implicit)
    e-mail networks, citation networks, co-author networks, terrorist collaboration networks

# SNA Research Topics

Universiteit Leiden

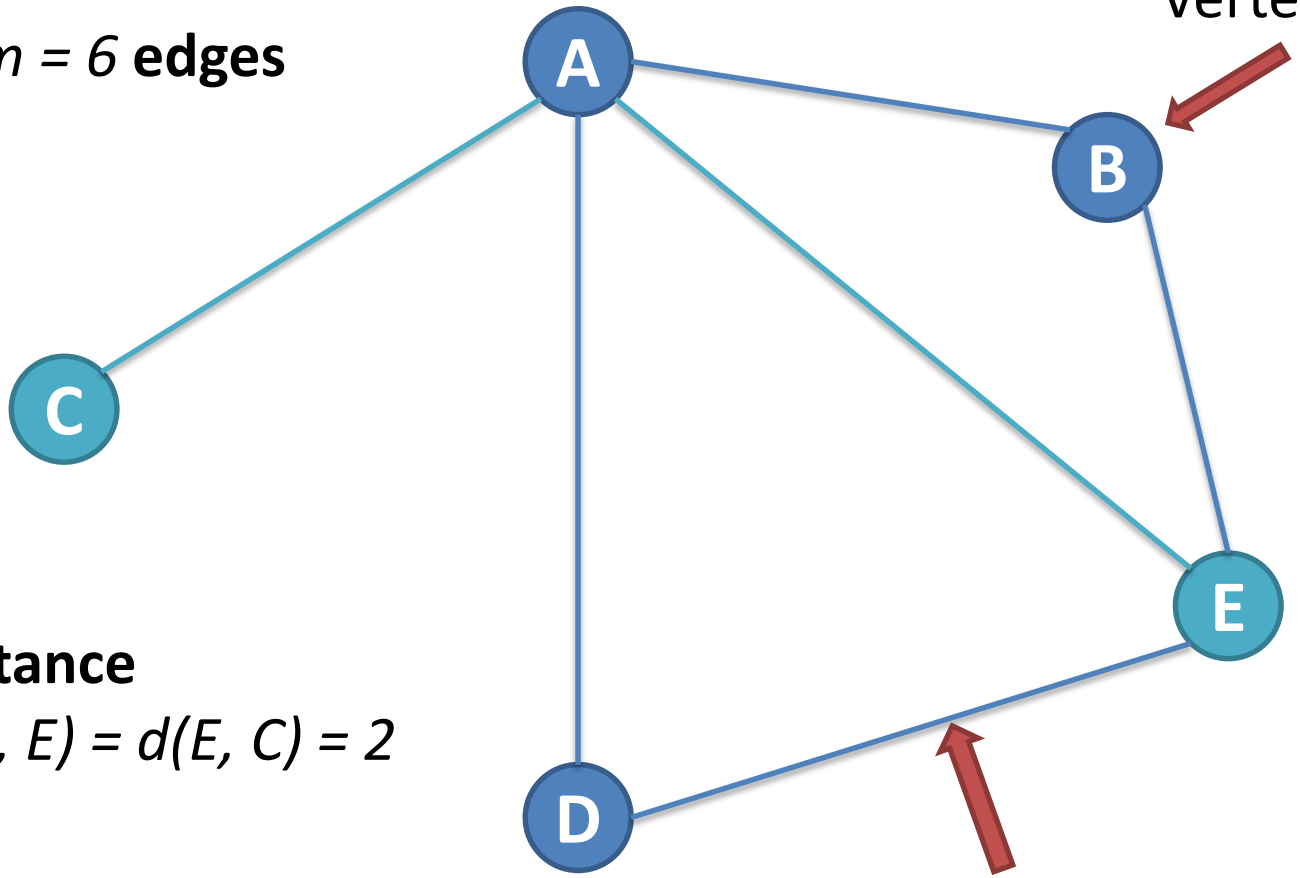
- Network analysis
  - Measuring
  - Modelling
  - Prediction
- Spread of Information
- Trust & Authority
- Community Detection
- Semantics
- Anonimity & Privacy

Universiteit Leiden

# Graph Mining

# Graphs

*n = 5* **nodes**
*m = 6* **edges**




**Distance**
*d(C, E) = d(E, C) = 2*

# Data / Graph Mining

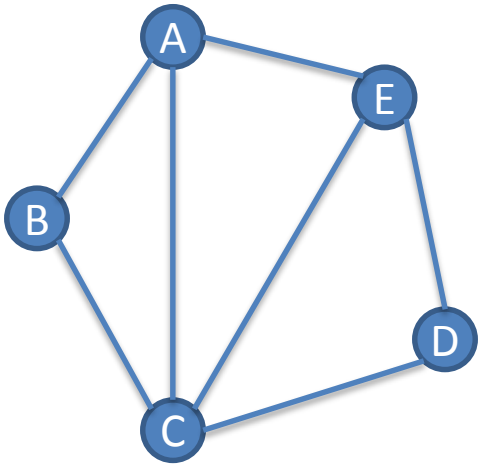
Universiteit Leiden

- **Data Mining**: looking for patterns in **Databases**

| id | name | date_of_birth | email |
|---|---|---|---|
| 23 | B. Stinson | **1983-08-15** | goforbarney@me.com |
| 42 | T. Mosby | **1983-08-15** | mosbydesigns@gmail.com |

classification, clustering, outlier detection, …

- **Graph Mining**: looking for patterns in **Graphs**




So what are we looking for?

# Graph Mining

Universiteit Leiden

- Graph models
- Centrality measures (local)
- Graph properties (global)

- Frequent subgraphs (pattern mining)
- Detecting cliques (clustering)
- Link prediction (classification)

- Various application domains

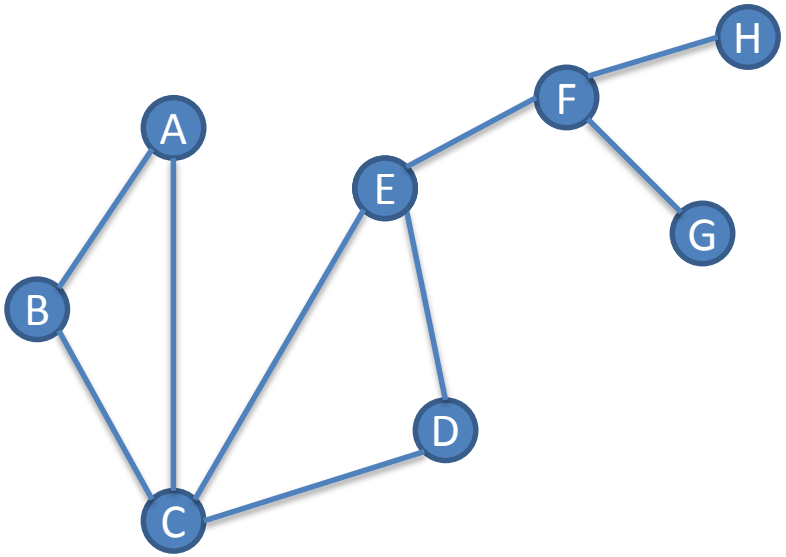
# Centrality measures

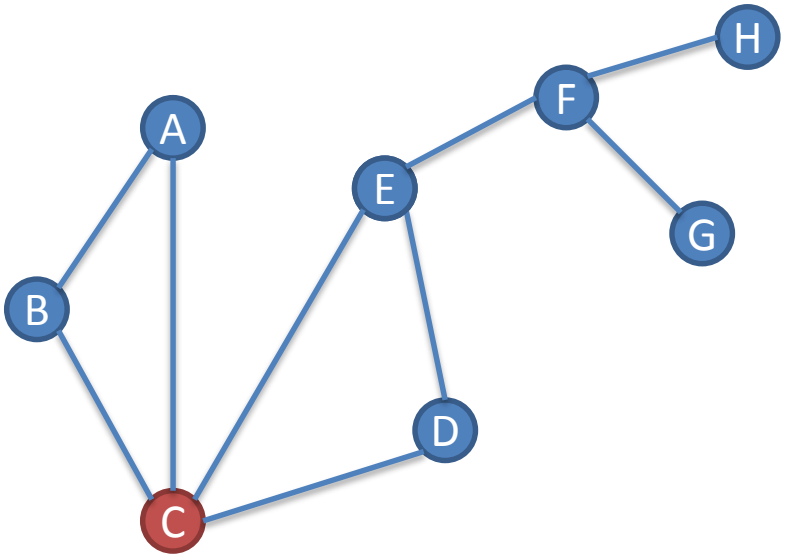
Universiteit Leiden

- Degree centrality
- Betweenness centrality
- Closeness centrality
- Graph centrality (eccentricity centrality)
- Eigenvector centrality
- Random walk centrality
- Hyperlink Induced Topic Search (HITS)
- PageRank

# Centrality

- Who has a central position in this graph?

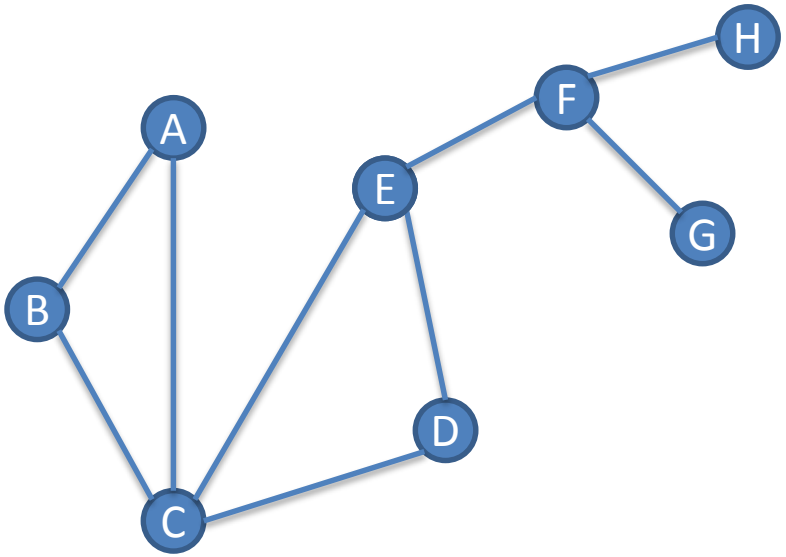


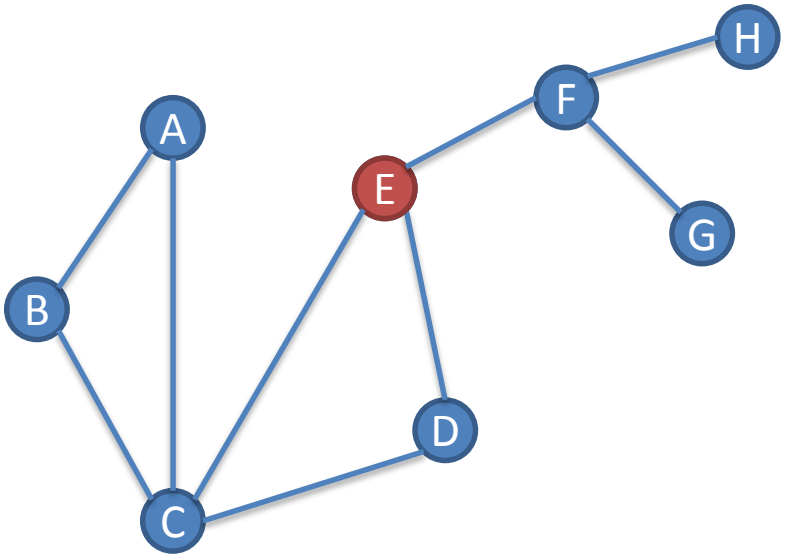




**Degree Centrality**:
*C* has the highest degree

# Centrality

- Who has a central position in this graph?

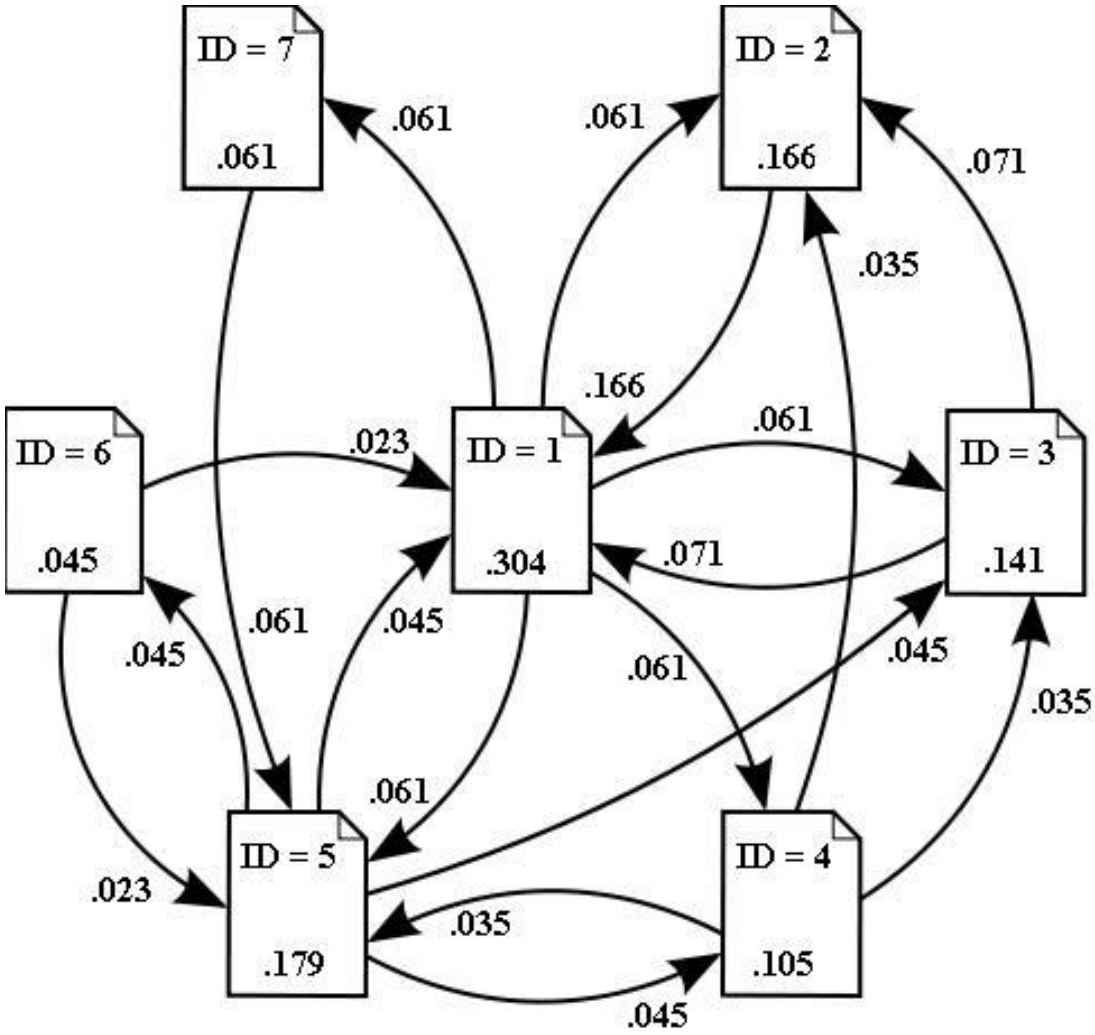



**Betweenness Centrality**:
*E* is part of the largest number of shortest paths

Universiteit Leiden

# Google PageRank



# Google PageRank

Universiteit Leiden

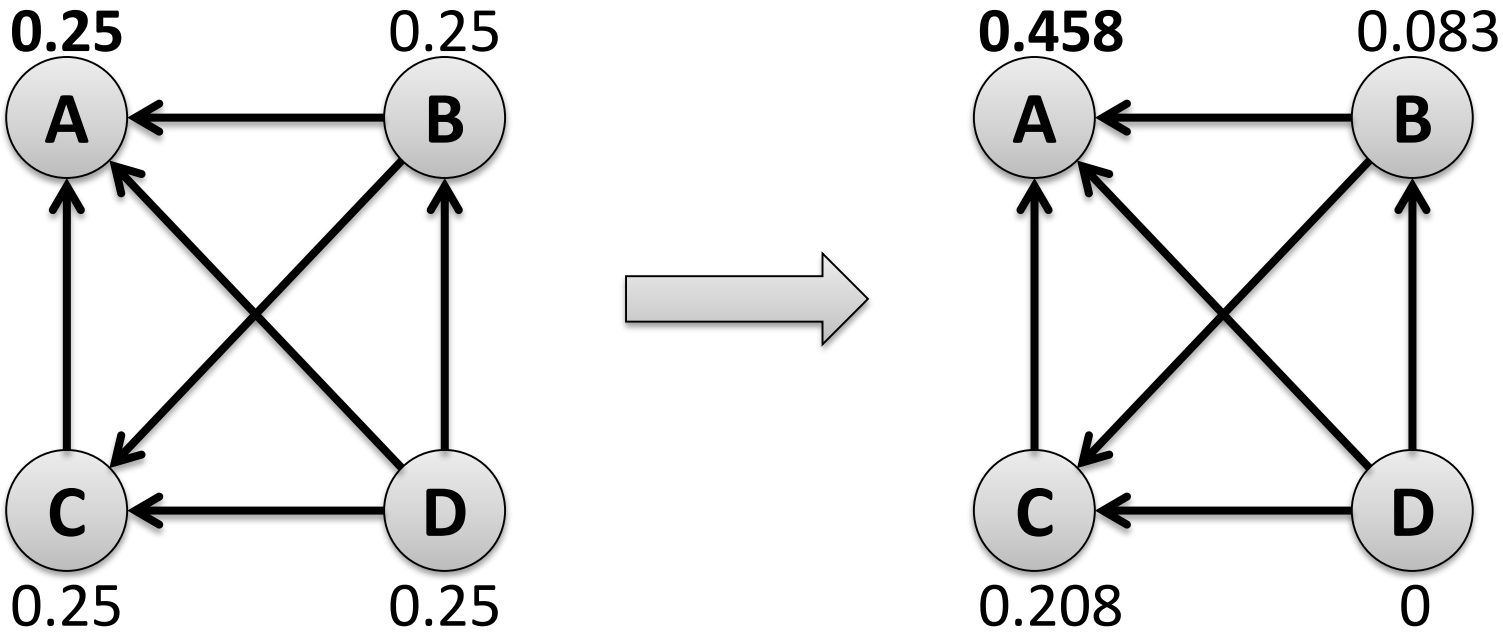
- 4 webpages *A*, *B*, *C* and *D* (*N = 4*)
- Initially: *PR(A) = PR(B) = PR(C) = PR(D) = 1/n*
- *L(A)* is the outdegree of page *A*
- Now if *B*, *C* and *D* each link to *A*, the simple PageRank *PR(A)* of a page *A* is equal to:

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

# Google PageRank

Universiteit Leiden

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$



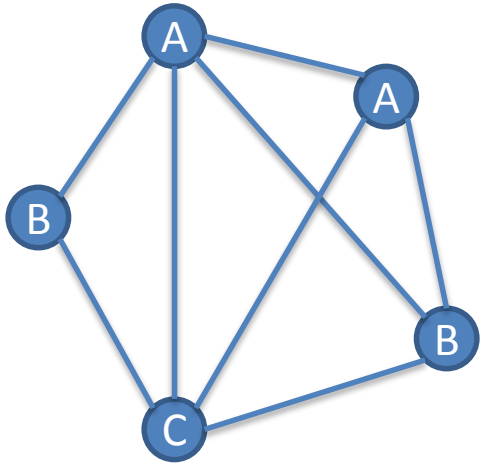
# Google PageRank

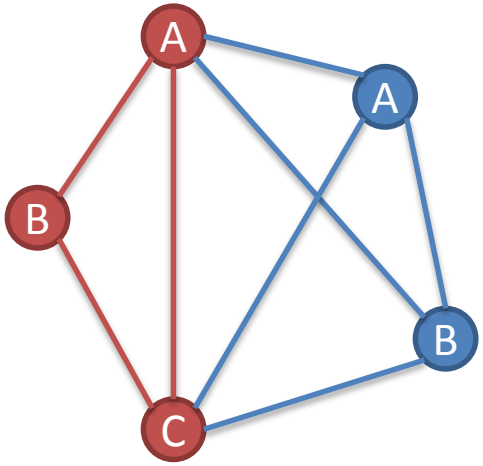
Universiteit Leiden

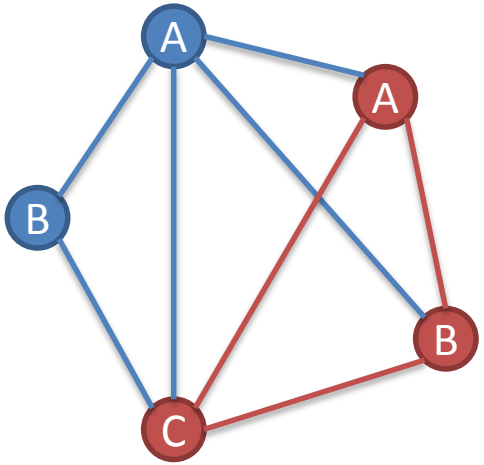
- **PageRank** as suggested by Larry Page in 1999
- $N$ = number of pages, $p_i$ and $p_j$ are pages
- $M(p_i)$ is the set of pages linking to $p_i$
- $L(p_j)$ is the outdegree of $p_j$
- $d = 0.85$, 85% chance to follow a link, 15% chance to jump to a random page (random surfer)
- $t = 0$ $$PR(p_i; 0) = \frac{1}{N}$$
- $t = t + 1$ $$PR(p_i; t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j; t)}{L(p_j)}$$

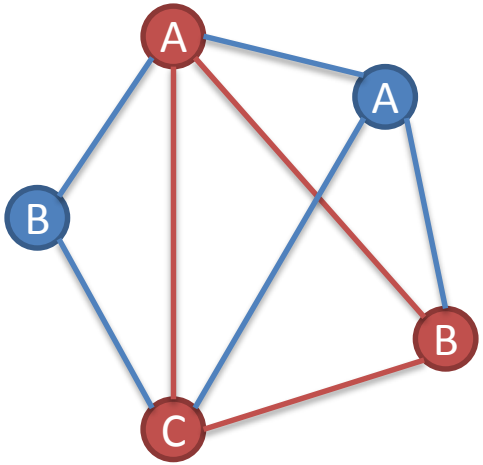
# Frequent Subgraphs

- What pattern occurs frequently in this graph?



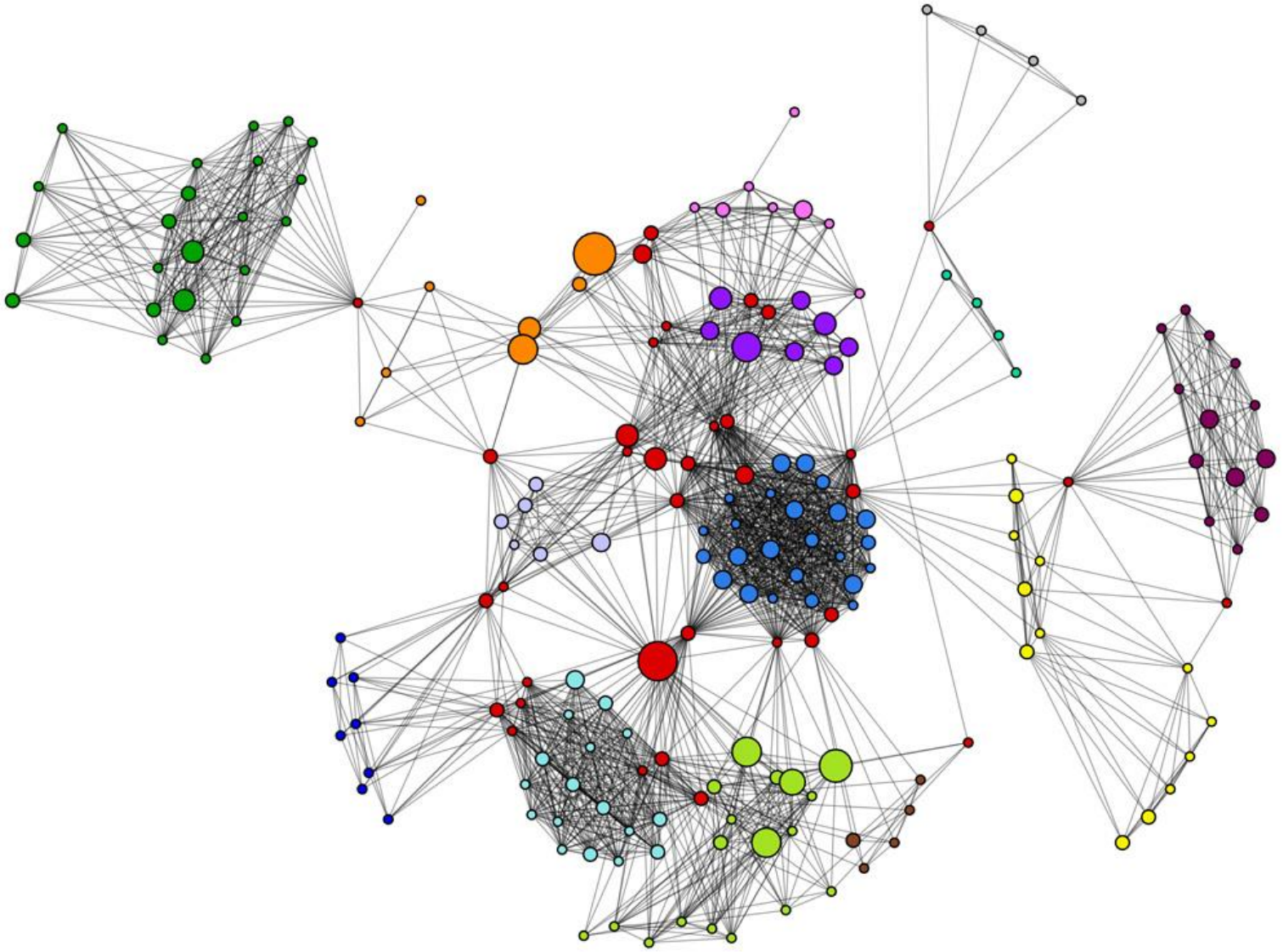




**Frequent Subgraph**: A-B-C

# Clustering



http://bayramannakov.wordpress.com

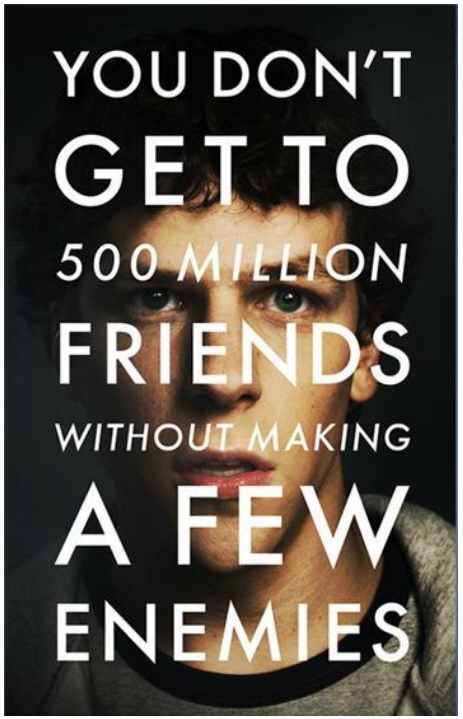
Universiteit Leiden


Spongecell
Jula
kiko
Trumba
eskobo
mayomi
Pageflakes
vimeo
skobee
shadows
gravee
YouTube
Zimbra
LookSmart
FURL | Your Personal Web File
smugmug
the ultimate in photo sharing
newsgator
Blogniscient
The bird's eye view of the Blogosphere
TiN FiNGER
shutterfly
feedia
PodDater
Feedster
favoor
Planzo.com
The Online Planning Community
ZAZZLE
Tailrank
TagWorld
nuvvo
Learn something new.
dogear
yakalike
grouper
ODDPOST
QOOP
iNods
Lulu
BR
BETA
blish
The easiest place to buy or sell digital content
facebook
FireAnt
simply hired
Veoh
theadcloud
rbloc.com
cafepress
Renkoo
pixagogo
gather
Agatra
browsr
oyogi
YEDDA
standpoint
meebo
EXTRA TASTY!
last.fm
Jotspot
Frappr!
jeteye
dabble
Writeboard
SHOUTWIRE
iKarma
KaNooble
AirSet
tech.memeorandum
CalendarHub
Suprglu
PIECING YOUR WEB TOGETHER
pando
zigtag
Findory
backfence
clipmarks
wayfaring
gOFFICE
AllPeers
orb
Rallypoint
Zoozio
blogbeat
Ziggs
zoto
vSocial
Boltfolio
wink
riya
Audible Wordcast
Opinity
Because your reputation matters!
reddit
measuremap
gumshoo
bluepulse
imvu
BETA
STREAMLOAD
Freedom for Your Digital Lifestyle
Ta-da Lists
FeedSky
jellyBarn.INC
nativetext
CONGOO
BETA
PODZINGER
BETA
RSS MAD
FeedTier
phanfare
twitter
PubSub
dPolls
flickr
Ning
Ookles
Strongspace
zoominfo
CASTPOST
ALPHA
WIKIPEDIA
The Free Encyclopedia
yubnub
AC
ASSOCIATEDCONTENT
BLOOP
BETA
ProjectSpaces
FeedBurner
Bloglines
purevolume.com
FOTOLOG
ourmedia
The Global Home for Grassroots Media
Yub.com
Spot Runner
BETA
myspace
a place for friends
NewsAlloy
Allmydata.com
gabbr.com
a social news community
blinkx
search
openomy
riffs
ajchat
alpha
jambo
NETWORKS
ROLLYO
BETA
ClipShack
beta
chatsum
Beta
PANDORA
looklater
beta
PROTOPAGE
FREE PERSONAL START PAGES
squishr
Webjay
PLAZES
Noodly
30
djigo
box


# Online Social Networks

# History

- 1997: SixDegrees.com
- 2000: Friendster
- 2003: LinkedIn & MySpace
- 2004: Hyves
- 2005: Facebook
- 2006: Twitter
- ……
- 2010: The Social Network (movie)



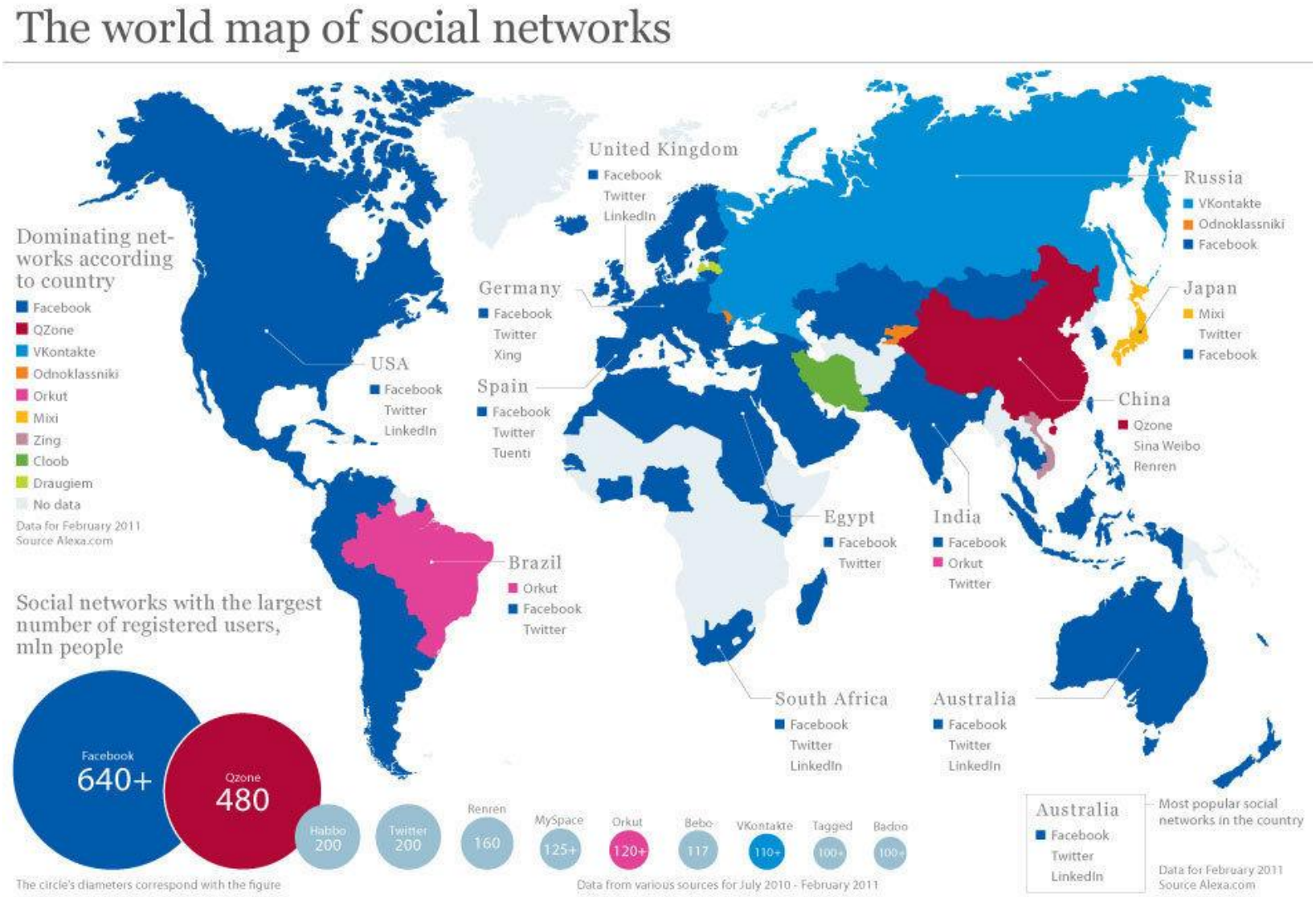
# Online Social Networks

Universiteit Leiden

- **User** (node) has a profile
- Profiles have **attributes** (labels/annotations)
- **Explicit links** (edges)
  - Social Links / Friendship links
  - User groups
- **Implicit links**
  - Social messaging
  - Common attributes
- Directed vs. undirected links

Universiteit Leiden

# 2011




Rumer Willis
Ashton Kutcher
LeVar Burton
Malik
GigaOM
Hodgman:
"... do you want to go candlepin bowling with me and Harrison Ford? ..."
Soleil Moon Frye ("Punky Brewster")
John Mayer
Starbucks
Brent Spiner (Star Trek's Data)
John Hodgman
Questlove (from The Roots)
Shaquille O'Neal
Lady GaGa
Rachel Maddow
Biz Stone (Twitter cofounder)
Robin Williams
Lance Armstrong
Fred Durst
Kevin Smith
Michael Arrington, TechCrunch blog
I Can Has Cheez-burger
Trent Reznor
Tony Hawk
Jenna Haze (Pornographic film star)
Dave Navarro
Barack Obama
Mandy Lynn (Playboy model)

# Example: Facebook

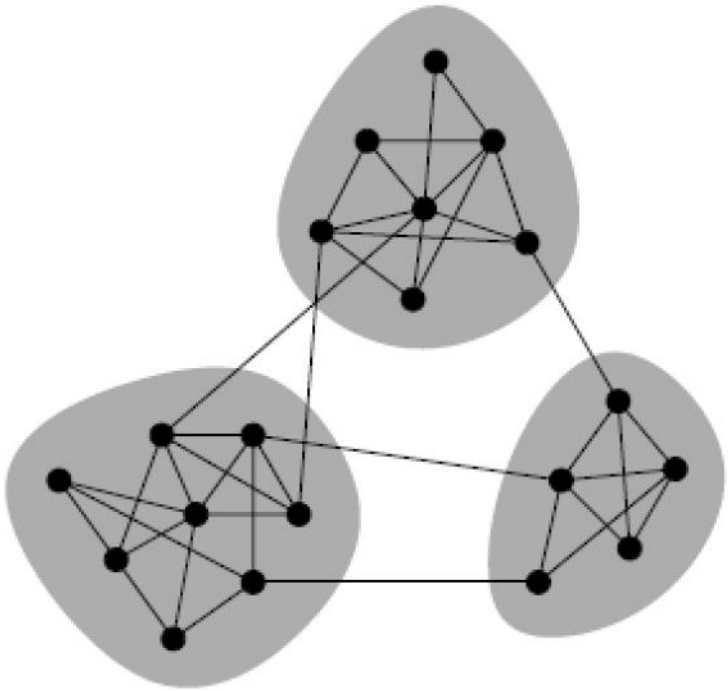
Universiteit Leiden

- More than **1 billion** active users
- Average user has 130 friends
- Estimated **100 billion** social links
- Over 600 million interactive objects (pages, groups and events)
- More than 45 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) shared each month

# OSNA Research Topics

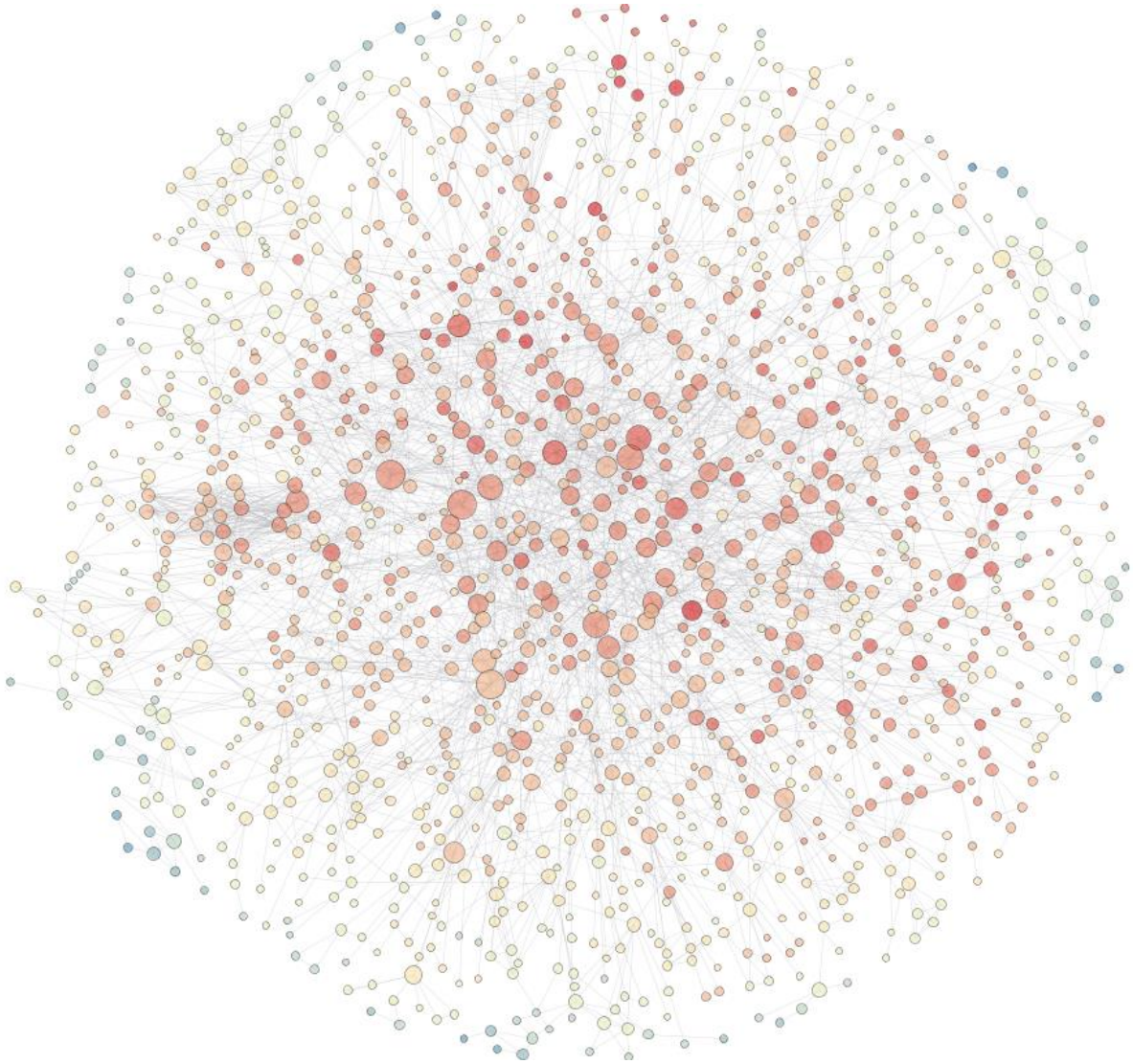
Universiteit Leiden

- User behavior
- Privacy & Anonymity
- Trust & Authorities
- Diffusion of information
- Sampling & Crawling
- Community Detection
- **Friendship Graph**




# Friendship Graph

# Friendship Graph



Universiteit Leiden

# Friendship Graph Analysis

Universiteit Leiden

- **Static analysis**
  - Densely connected core
  - Fringe of low-degree nodes
  - Few isolated communities & singletons
  - **Static properties**
    - Node **degree distribution**, average distance, diameter
    - Edge/node ratio, level of symmetry
    - Number of cliques, *k*-cliques, etc.
    - **Small world phenomanon**

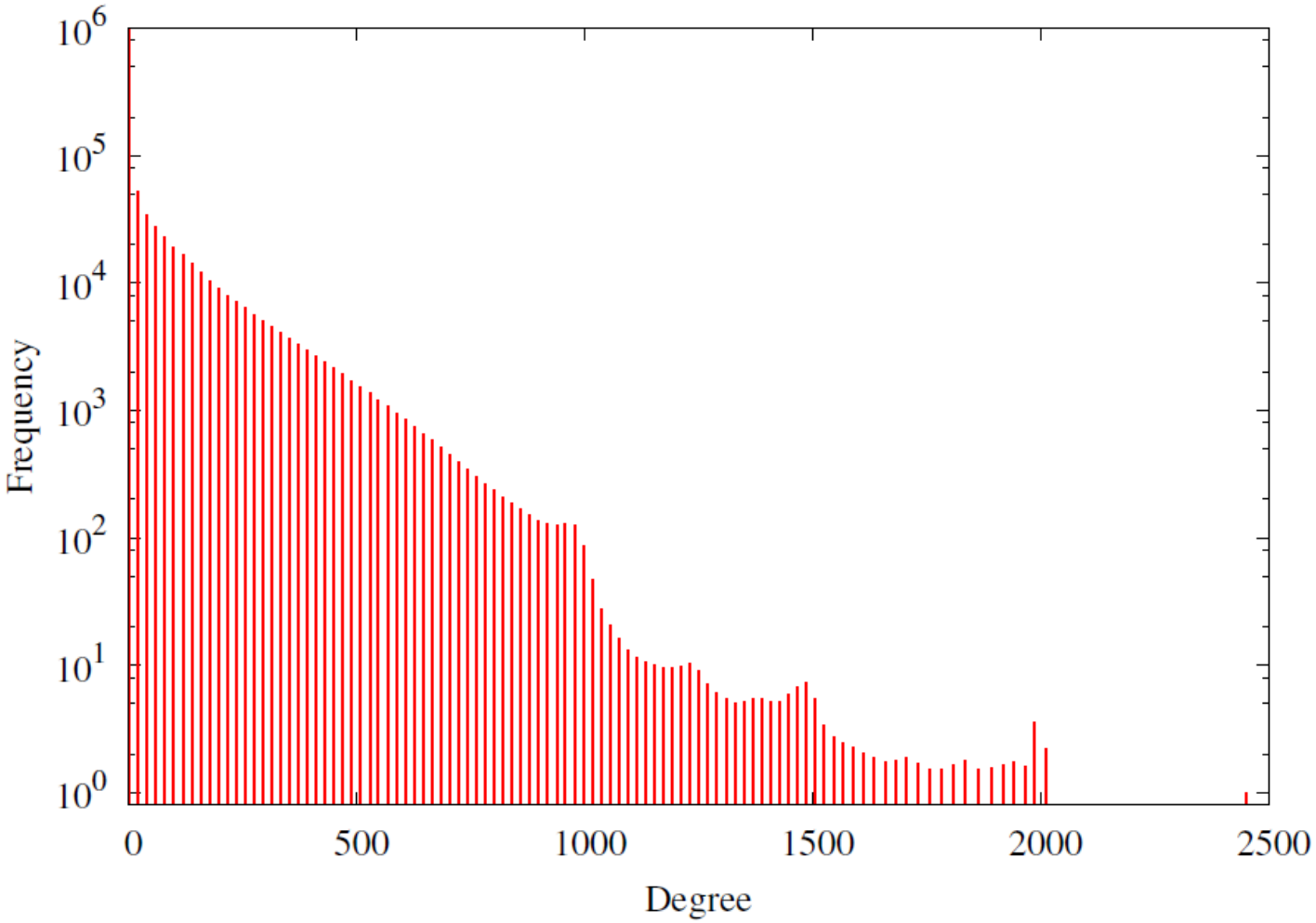
# Static properties

Universiteit Leiden

| Dataset | Nodes | Links | Average Degree | Average Distance | $\Delta$ |
|---|---|---|---|---|---|
| AstroPhys [12] | 17,903 | 396K | 21 | 4.15 | 14 |
| Enron [9] | 33,696 | 362K | 10 | 4.07 | **13** |
| Flickr [15] | 1,624,992 | 30.9M | 18 | 5.38 | 24 |
| Hyves | 8,057,981 | 871M | 112 | 4.75 | 25 |
| LiveJournal [15] | 5,189,809 | 97.4M | 19 | 5.48 | 23 |
| Orkut [15] | 3,072,441 | 234M | 76 | 4.16 | 10 |
| Skitter [11] | 1,696,415 | 22.2M | 13 | 5.08 | **31** |
| YouTube [15] | 1,134,890 | 5.98M | 5.3 | 5.32 | 24 |
| Web [13] | 855,802 | 8.64M | 10 | 6.30 | 24 |
| Wikipedia [5] | 2,213,236 | 23.5M | 11 | 4.81 | 18 |

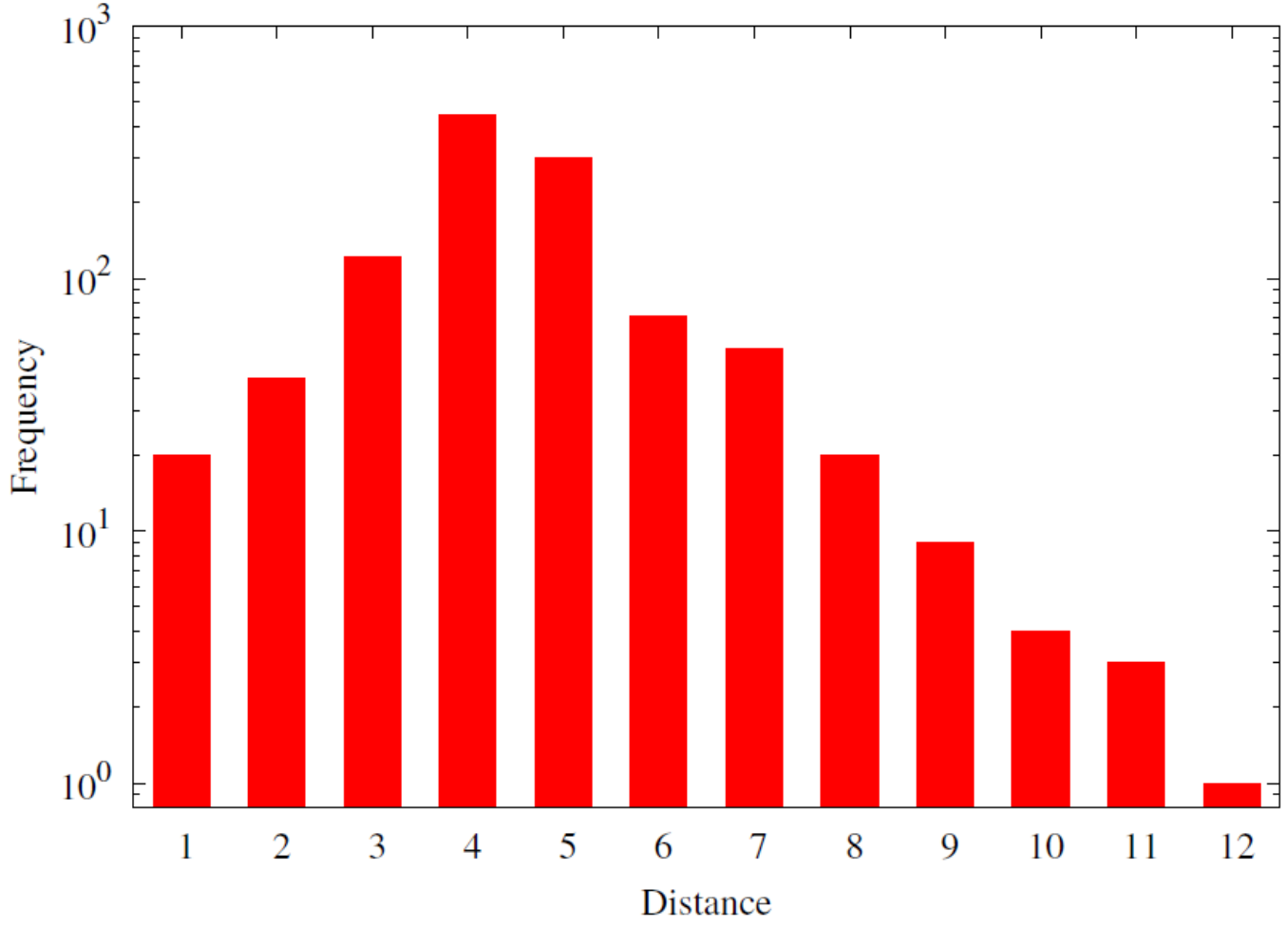
# Degree Distribution

Universiteit Leiden



# Distance Distribution

Universiteit Leiden



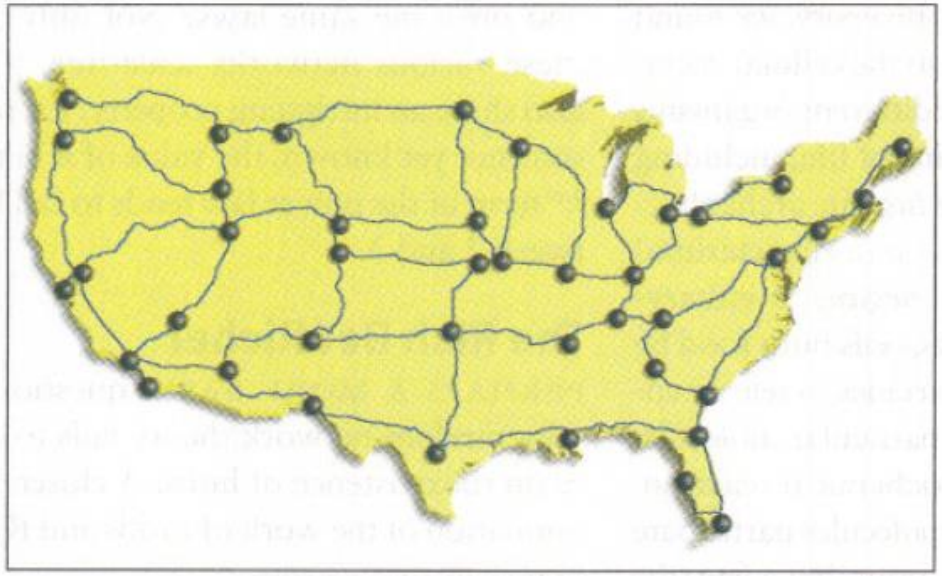
# Small World Networks

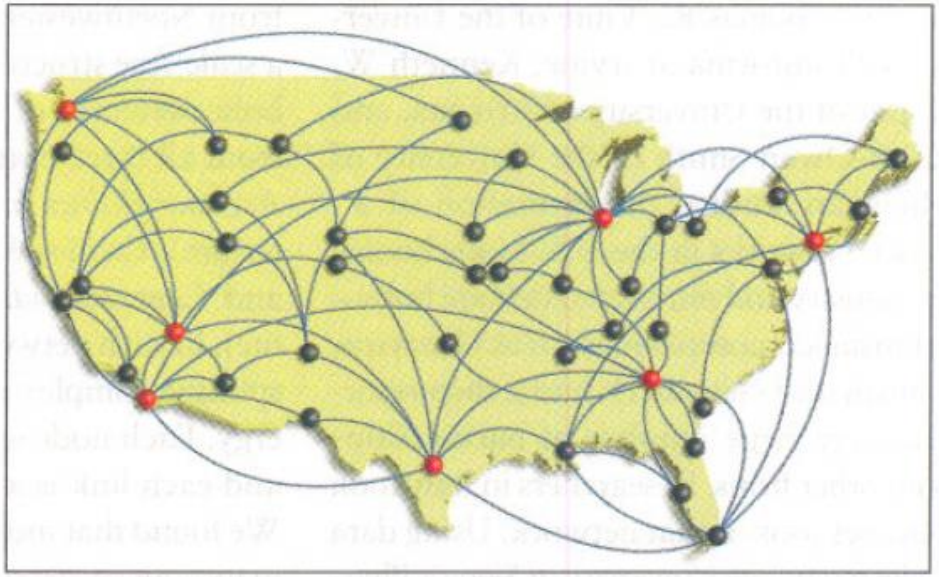
Universiteit Leiden

- Class of networks with certain properties:
  - Sparse graphs
  - Highly connected
  - Short average node-to-node distance: $d \sim log(n)$
  - Fat tailed power law node degree distribution
  - Densely connected core with many (near-)cliques
  - Existence of hubs: nodes with a very high degree
  - Fringe of low(er)-degree nodes

# Regular vs. Small World

Universiteit Leiden





# Small World Networks

Universiteit Leiden

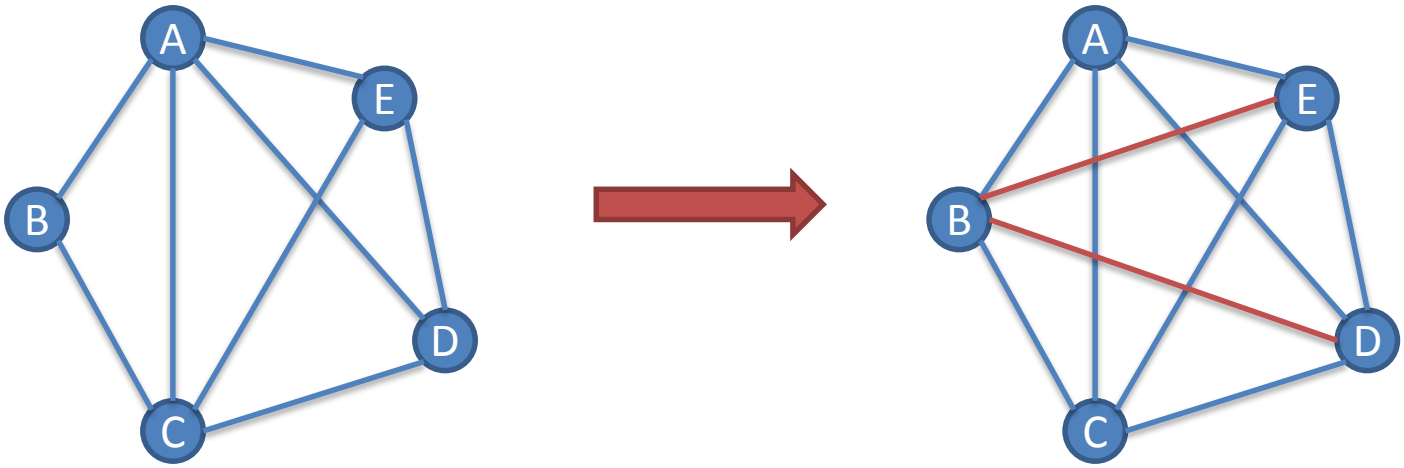
- Other examples of small world networks
  - Web graphs
  - Gene networks
  - E-mail networks
  - Telephone call graphs
  - Information networks
  - Internet topology networks
  - Scientific co-authorship networks
  - Corporate networks (interlocks or ownerships)

# Friendship Graph Analysis

Universiteit Leiden

- Static analysis
- **Dynamic analysis**
  - Network evolution
  - Network modelling
  - Network growth
  - **Link Prediction**
    - **Triadic Closure**
    - **Preferential Attachment**
  - **Semantic Link Prediction**

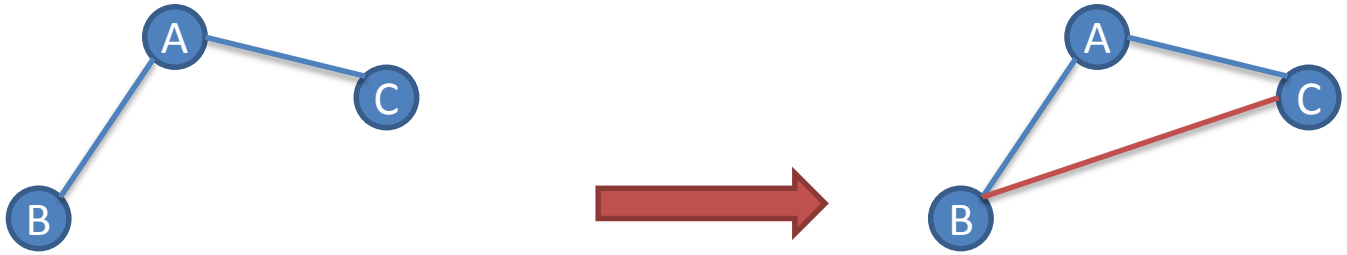
# Link Prediction




- Two principles: preferential attachment and triadic closure

# Triadic Closure

Universiteit Leiden

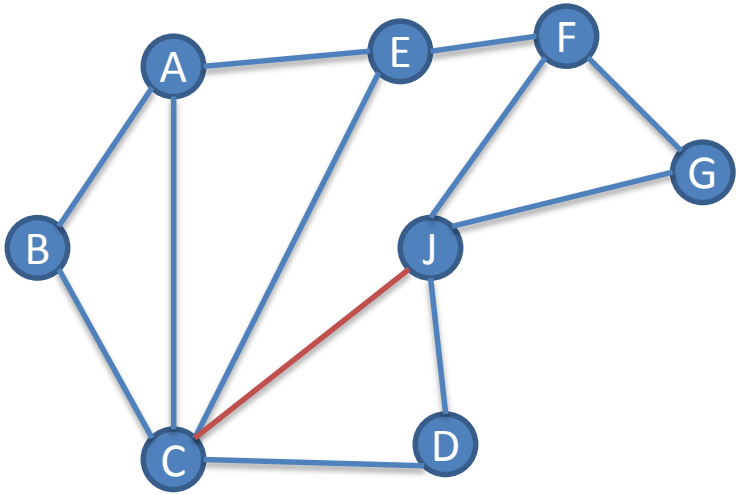



- Two principles: preferential attachment and triadic closure

# Preferential Attachment

Universiteit Leiden

- Nodes with a large degree acquire new links at a faster rate.

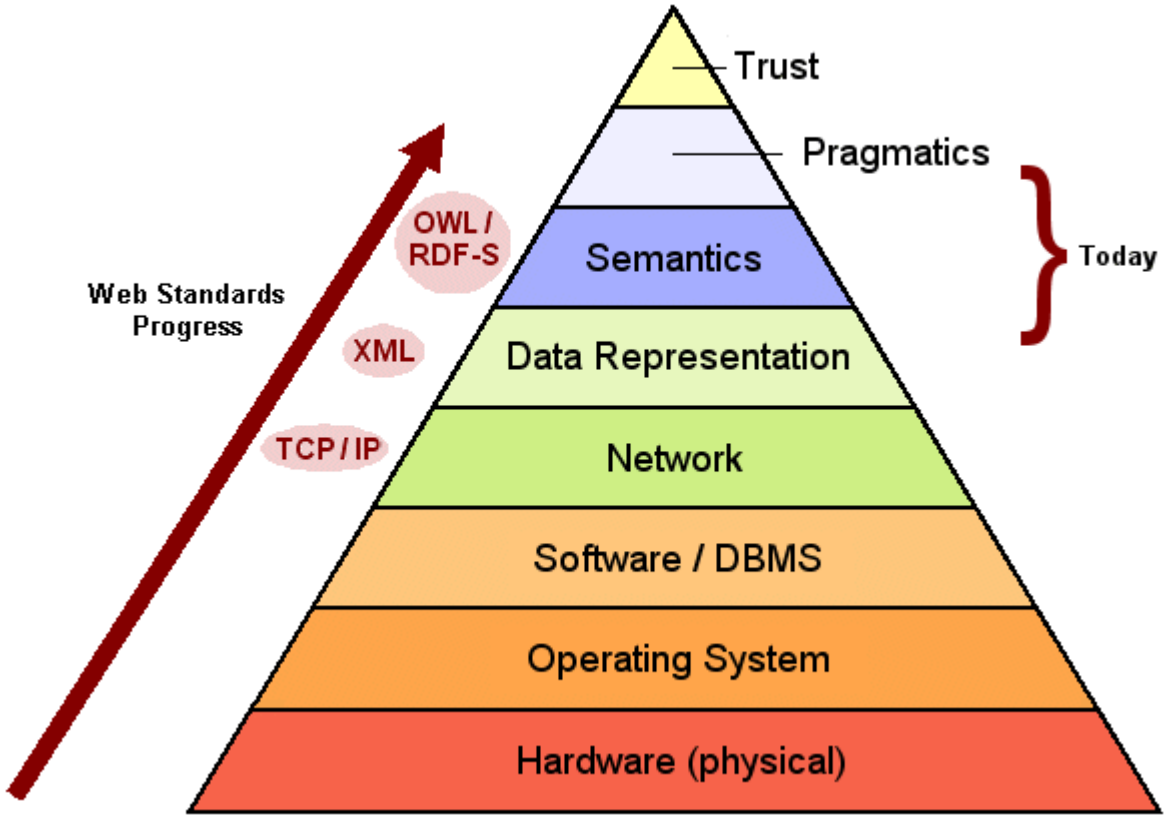



# Semantics

# "Call for Semantics"

Universiteit Leiden

- Structured data volumes grow rapidly
- "Semantic Web"



# "Call for Semantics"

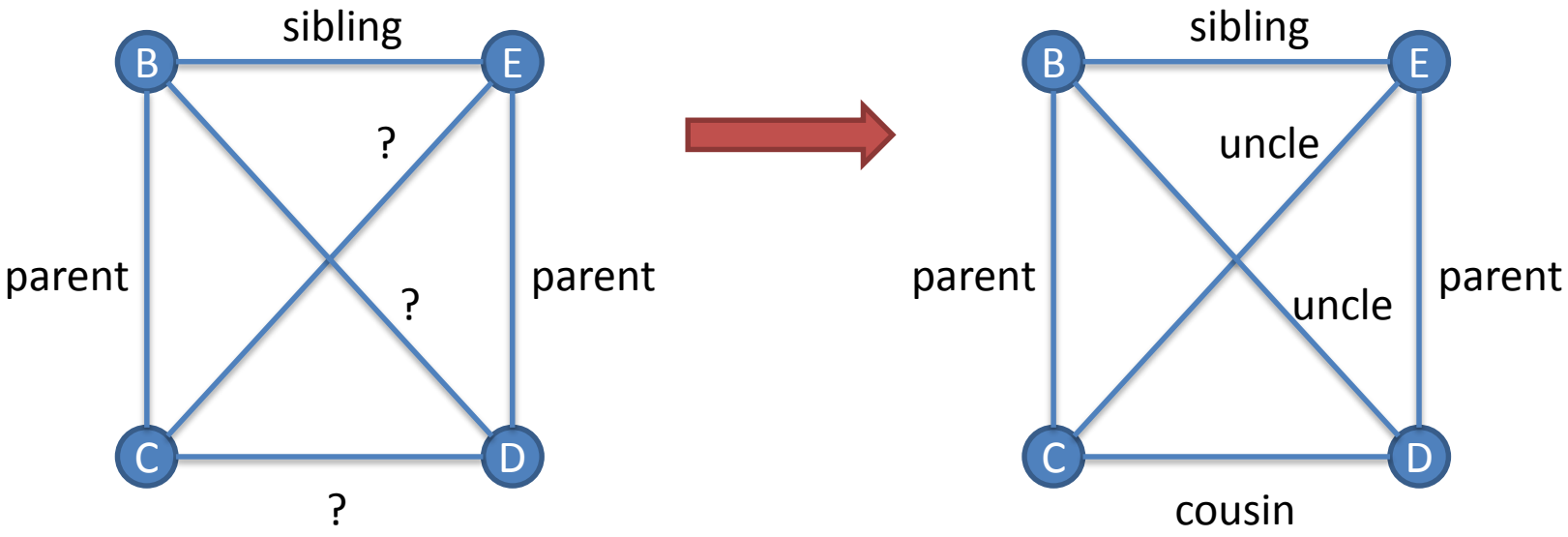
Universiteit Leiden

- **What is the meaning of a link?**
- What type of relation is defined by a link?
- Are there any **wrong** links?
- What is the **strength** of a link?
- Are link descriptions missing?

# Semantic Link Prediction

Universiteit Leiden

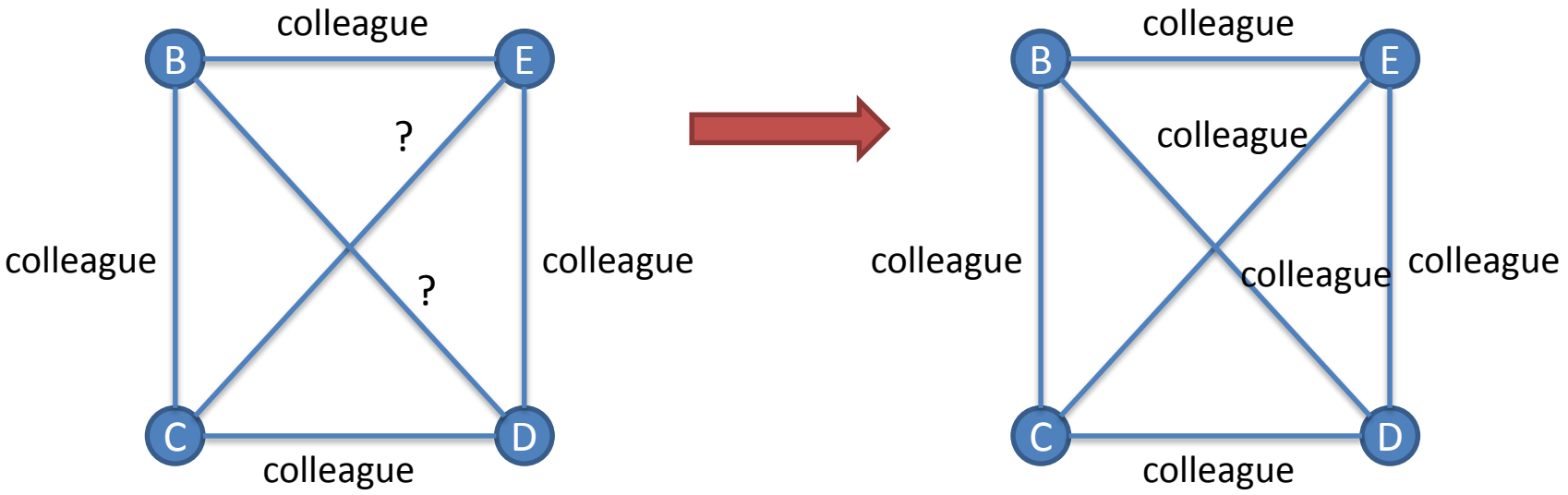
- Deterministic



# Semantic Link Prediction

Universiteit Leiden

- Predictive (learning, classification, etc.)



# Conclusions

Universiteit Leiden

- Online social networks are an **excellent domain of study** for data (or graph-) miners
- Social Network Analysis is important for **many areas of research**, not only computer science
- **Semantics** within large networks are becoming increasingly more important
- Challenges may be found in the **temporal** (dynamic) **analysis** of social networks