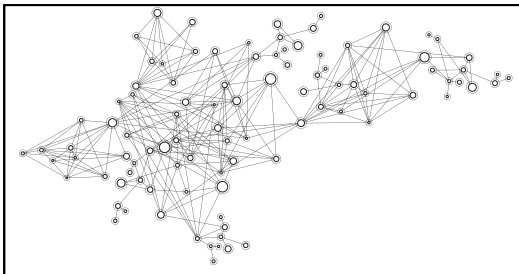


## Network science in biology

measuring, visualizing and modelling real world complex networks

Petra Vertes

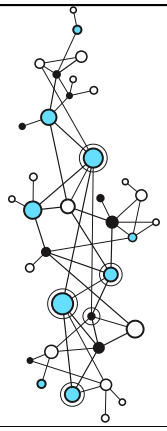
## Before we begin:

1. Register
   A. Subject and level of studies
2. Link to last lecture's slides:
   http://tinyurl.com/od652m2
3. Will you be writing an essay for this course?
   A. Don't know yet
   B. Yes
   C. No

## Before we begin:

"…a perfect analogy could be drawn between the mathematics of the network and the mathematics of a Bose gas if each node in the network were thought of as an energy level, and each link as a particle. These results have implications for any real situation involving random graphs, including the world wide web, social networks, and financial markets."

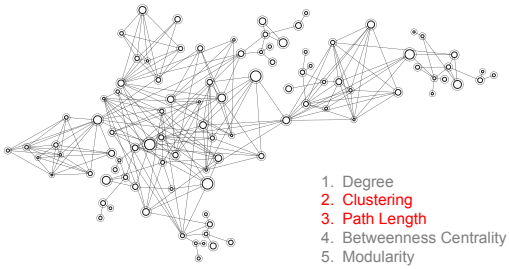
http://en.wikipedia.org/wiki/Bose–Einstein_condensation_(network_theory)



## Overview of topics

1. What is a network? – examples from social and biological sciences.
2. Constructing and representing complex networks.
3. Topological properties of networks – how to measure them and why they matter?
4. Network analysis in biological sciences – six examples
5. Generative modelling of networks – why and how?
6. Getting hold of data and code – tools and resources for network analysis

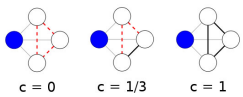
## 3. Network Measures (and why they matter)



1. Degree
2. Clustering
3. Path Length
4. Betweenness Centrality
5. Modularity
6. Motifs

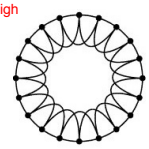
## Clustering Coefficient of a Node

$$C_i = \frac{\textit{Links between 'friends'}}{\textit{Total possible number}}$$

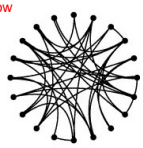


c = 0 c = 1/3 c = 1

High C



Low C



# Weighted measures of Clustering

PHYSICAL REVIEW E 75, 027105 (2007)

**Generalizations of the clustering coefficient to weighted complex networks**

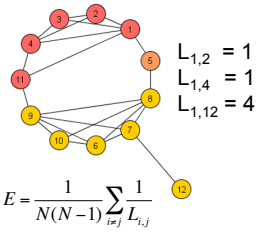
Jari Saramäki,[1,*] Mikko Kivelä,[1] Jukka-Pekka Onnela,[1,2] Kimmo Kaski,[1] and János Kertész[1,3]
[1]*Laboratory of Computational Engineering, Helsinki University of Technology, P.O. Box 9203, FIN-02015 HUT, Finland*
[2]*Department of Physics, Clarendon Laboratory, University of Oxford, Oxford, OX1 3PU, United Kingdom*
[3]*Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary*
(Received 31 August 2006; published 23 February 2007)

The recent high level of interest in weighted complex networks gives rise to a need to develop new measures and to generalize existing ones to take the weights of links into account. Here we focus on various generalizations of the clustering coefficient, which is one of the central characteristics in the complex network theory. We present a comparative study of the several suggestions introduced in the literature, and point out their advantages and limitations. The concepts are illustrated by simple examples as well as by empirical data of the world trade and weighted coauthorship networks.

DOI: 10.1103/PhysRevE.75.027105 PACS number(s): 89.75.Hc, 87.16.Ac, 89.65.−s

# Path Length (or Efficiency)

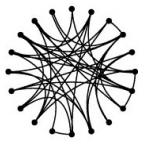
$L_{i,j}$ = *Shortest path between nodes i and j*



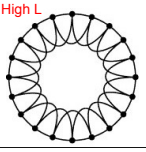

$L_{1,2} = 1$
$L_{1,4} = 1$
$L_{1,12} = 4$

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{L_{i,j}}$$

Low L



High L

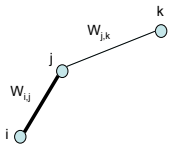


# Weighted Efficiency

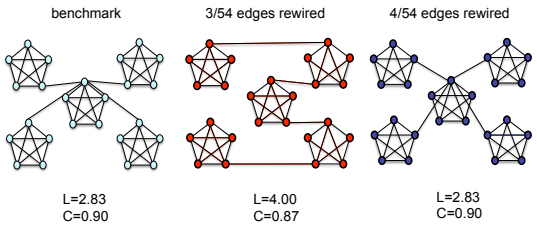
$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{L_{i,j}}$$

$$E_{weigthed} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{i,j}}$$

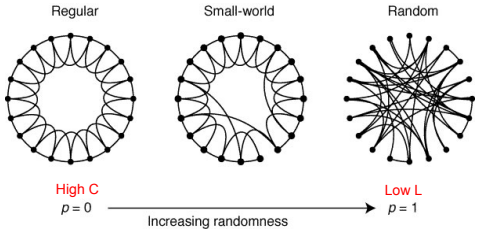
Where $d_{i,j}$ is the weighted version of the path length such as the sum of the inverse weight $w_{i,j}$ of each edge (i to j) in the path.




# Comparing Networks

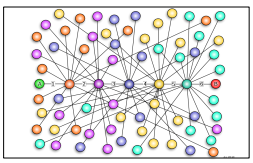



# Small- World Networks: Why it is Fun to Surf the Web




*Watts & Strogatz (1998) Nature*

# Milgram's small-world experiment



**1967, Harvard:**

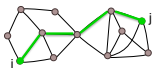
Milgram sent hundreds of letters to people in Nebraska asking them to forward it to personal acquaintances who might be able to bring it closer to the target person: a stockbroker in Boston.

A significant problem was that often people refused to pass the letter forward. However, among the letters that did reach their target, the average path length fell around five and a half or six.

There are a number of methodological critiques of the Milgram Experiment, which suggest that the average path length might actually be smaller or larger than Milgram expected – can you think of some?

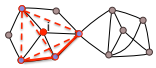
## Quizz

Path Length (Efficiency)



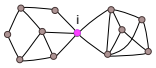

$L_{i,j} = ?$

Clustering



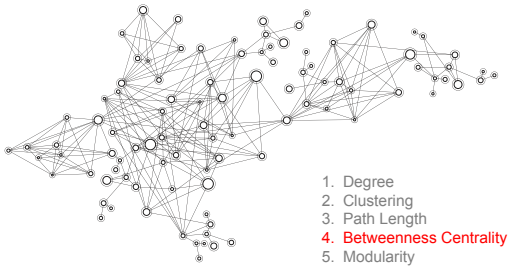
$C_i = ?$

Degree (Hubs)



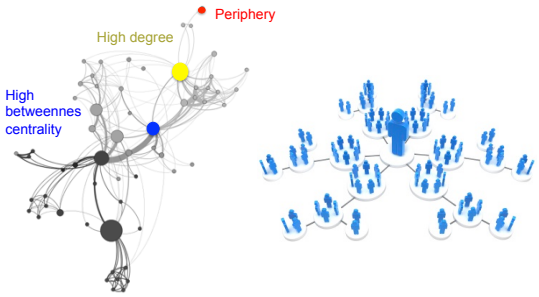

$d_i = ?$

## 3. Network Measures (and why they matter)

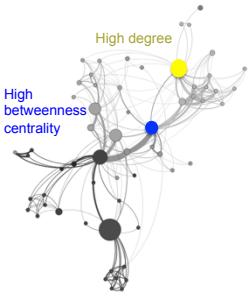


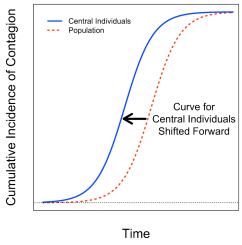
1. Degree
2. Clustering
3. Path Length
4. Betweenness Centrality
5. Modularity
6. Motifs

## Centrality Measures Identify Key Players

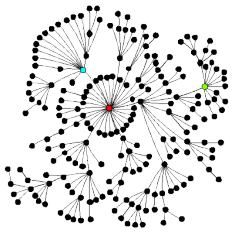



## Central individuals are Early Adopters



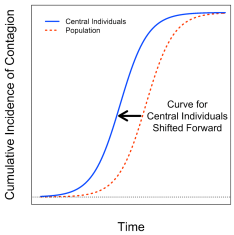

## Finding Central Individuals: The Friendship Paradox

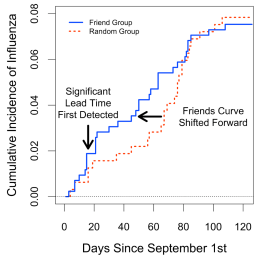


On average, the friends of randomly selected people possess more links (have higher degree) and are also more central (e.g., as measured by betweenness centrality) to the network than the initial, randomly selected people who named them.

*Scott L. Feld in 1991*

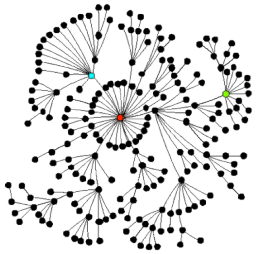
## Early Detection of an Epidemic



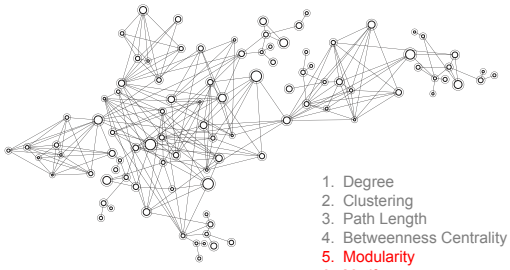

*Christakis & Fowler, 2010*

# Immunization of central individuals?



- Diffusion theories predict a particular threshold for the propagation of contagion throughout a population: any virus that is less infectious than a given threshold will die out.
- In scale-free networks this threshold is zero. Because hubs have so many links they are likely to get infected early and will then pass it on to huge numbers of others.
- Is the traditional public health approach of random immunization wise? Could we target hubs?

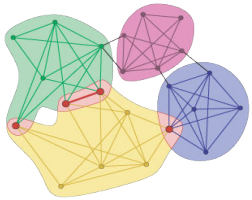
# 3. Network Measures (and why they matter)



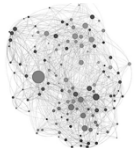
1. Degree
2. Clustering
3. Path Length
4. Betweenness Centrality
5. Modularity
6. Motifs

# Modularity

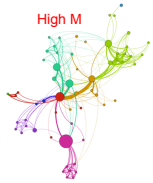
*M* = *Fraction of within-module connections compared to what is expected by chance (for the optimal partition).*



Low M



High M



# Network motifs

*Network motifs* are subgraphs of a few nodes which appear in directed networks more often than would be expected by chance.

*Image: Milo et al., Science 303, 1538 (2004)*

# Superfamilies

Alon (2004) showed that the frequency signatures of network motifs classify networks into *superfamilies*.

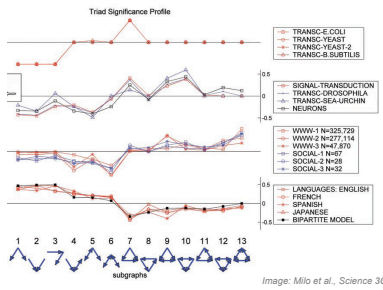



*Image: Milo et al., Science 303, 1538 (2004)*

# 4.

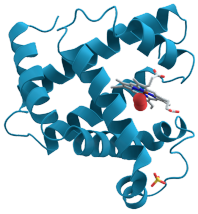
## Networks in Biology

(Cell biology 101)

1. Protein-protein interaction networks
2. Epistasis interaction networks
3. Drug target networks
4. The diseasome
5. Coexpression networks
6. Brain networks

# Cell biology 101 - Proteins

Proteins are long molecules made up of a chain of amino acids (smaller molecules). The sequence of these amino acids determines it folding into a specific three-dimensional structure and its activity.

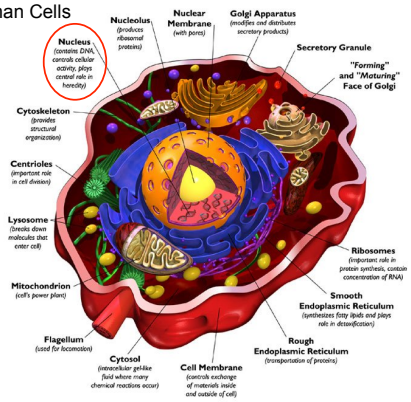


Proteins perform a vast array of functions within living organisms, including mechanical support, response to stimuli, catalyzing metabolic reactions (enzymes), storing and transporting molecules from one location to another, facilitating the immune response (antibodies).

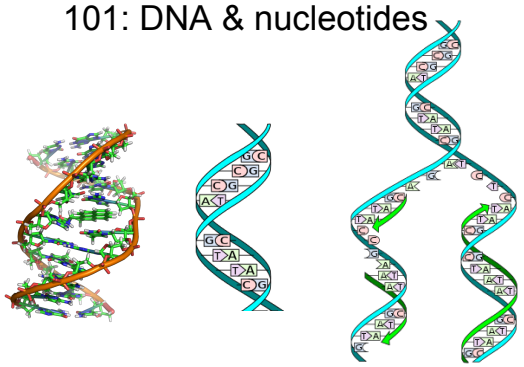
Different cells 'express' different types of proteins in different amounts at different times.

Examples:
- Keratin – key material in skin, hair, nails
- Hemoglobin – oxygen transporting protein in red blood cells
- P53 – regulates cell cycle
- Insulin – hormone controlling glucose absorption (from blood) by muscle and fatty tissues

# 101: Human Cells

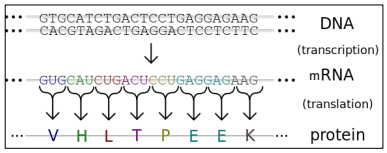



# 101: DNA & nucleotides



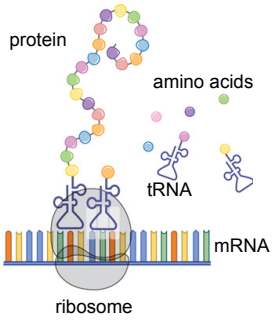
# 101: The genetic code

The genetic code is the set of rules by which information encoded within genetic material (DNA) is translated into proteins by living cells. The code defines how sequences of nucleotide triplets, called *codons*, specify which amino acid will be added next during protein synthesis.




While the genetic code determines the protein sequence for a given coding region, other genomic regions can influence when and where these proteins are produced.

# 101: Translation

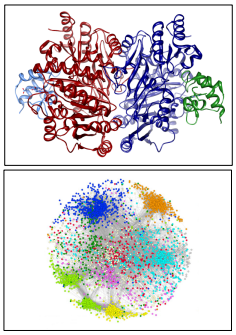



# 4.

## Networks in Biology

(Cell biology 101)
1. Protein-protein interaction networks
2. Epistasis interaction networks
3. The diseasome
4. Drug target networks
5. Coexpression networks
6. Brain networks

## Protein-protein interaction networks



a.k.a: **The Interactome**

**Nodes:** Proteins

**Links:**
Protein–protein interactions (PPIs) refer to intentional physical contacts between two or more proteins as a result of biochemical events and/or electrostatic forces.

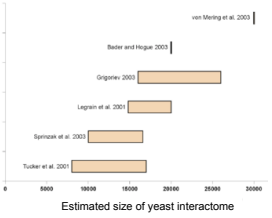
## Saccharomyces cerevisiae (yeast)

S. cerevisiae has developed as a model organism in cell biology because:

- As a single-cell organism, S. cerevisiae is small with a short generation time (doubling time 1.25–2 hours at 30 °C) and can be easily cultured.
- S. cerevisiae is easily amenable to genetic manipulations (addition and deletion of genes).
- As a eukaryote, S. cerevisiae shares the complex internal cell structure of plants and animals. Many proteins important in human biology were first discovered by studying their homologs in yeast; these proteins include cell cycle proteins, signaling proteins, and protein-processing enzymes.

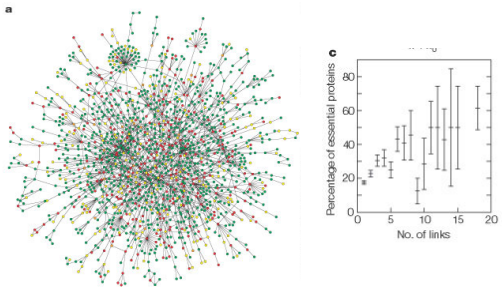
## Size of the interactome

- It has been suggested that the size of an organism's interactome correlates well with the biological complexity of the organism.
- Protein–protein interaction maps containing several thousands of interactions are now available for several organisms (viruses, bacteria, yeast, C.elegans, Drosophila, Human) - but none of them is presently complete.
- The yeast interactome has been estimated to contain between 10,000 and 30,000 interactions. Larger estimates often include indirect or **predicted interactions**.




Estimated size of yeast interactome

## Lethality and centrality



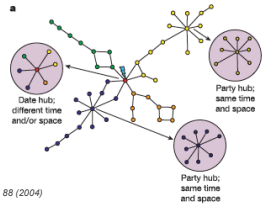

Protein-protein interaction network of yeast is scale-free. Yeast knockouts of yeast genes encoding hubs are approximately threefold more likely to confer lethality than those of non-hubs.

## Protein-protein networks

We can distinguish two types of hubs in protein-protein interaction networks:

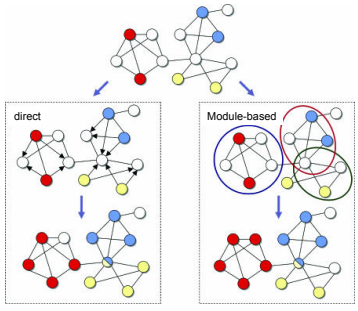
- **Party hubs**, which interact with several other proteins *simultaneously*.
- **Date hubs**, which interact with several other proteins *sequentially*.

Date hubs organize the proteome, connecting biological processes - or modules - to each other, whereas party hubs function inside modules.




*Han et al., Nature **430**, 88 (2004)*

## Network-based predictions of protein function



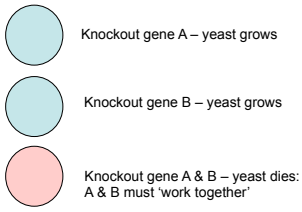

*Sharan et al (2007) Mol Syst Biol*

# 4.

## Networks in Biology

(Cell biology 101)
1. Protein-protein interaction networks
2. Epistasis interaction networks
3. The diseasome
4. Drug target networks
5. Coexpression networks
6. Brain networks

## Epistasis interaction networks




Global mapping of the yeast gene interaction network: 291 genetic interactions amongst 204 yeast genes

*Hin Yan Tong, Science 294: 2364 (2001)*

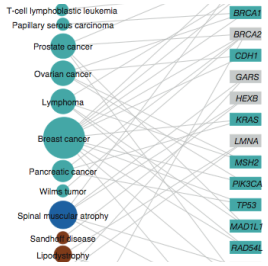
# 4.

## Networks in Biology

(Cell biology 101)
1. Protein-protein interaction networks
2. Epistasis interaction networks
3. The diseasome
4. Drug target networks
5. Coexpression networks
6. Brain networks

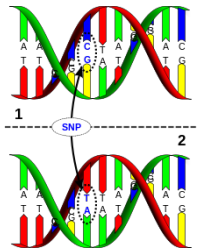
## The diseasome

The bipartite network of *diseases* and *disease-related genes* is also known as the *diseasome*.




*Goh et al. PNAS **104**, 8685 (2007).*

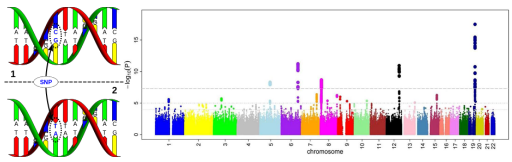
## Single Nucleotide Polymorphisms (SNPs)




- A SNP is a DNA sequence variation occurring commonly within a population.
- These genetic variations underlie differences in our susceptibility to disease.
- For example, a SNP in the APOE gene is associated with a higher risk for Alzheimer disease.

## Genome-wide association studies - GWAS

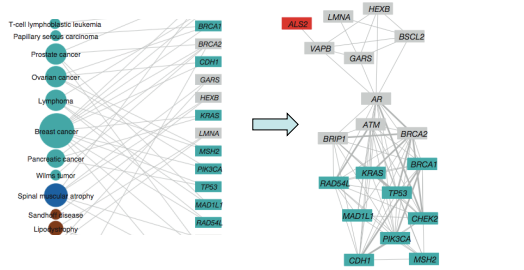
A genome-wide association study (GWAS) is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases.




An illustration of a Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. In this example the peaks indicates genetic variants that are more often found in individuals with constrictions in small blood vessels.

# Disease Gene Network

The *projection* of the diseasome onto disease genes, creates a *Disease Gene Network* (DGN).




*Goh et al. PNAS 104, 8685 (2007).*