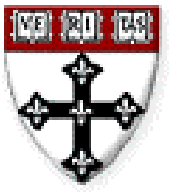
BIOCONDUCTOR
open source software for bioinformatics

Graphs,EDA and Computational Biology

Robert Gentleman

rgentlem@hsph.harvard.edu

www.bioconductor.org



Outline

- General comments
- Software
- Biology
- EDA
- Bipartite Graphs and Affiliation Networks
- PPI and transcription

General Comments

- in this talk I will outline some open problems rather than give solutions to them
- graphs are a rich data structure and it seems that there will be many interesting statistical challenges associated with them
- these will be both mathematical and computational

General Comments

- perhaps the biggest lesson to be learned here is to be careful to interpret the data correctly
- not all graphs are the same
- pair-wise information is different from whole-set information

General Comments

- in statistical research social network analysis and graphical models are the two areas that have historically used graphs
- *Social Network Analysis*, Wasserman and Faust, is a good reference
- for graphical models the books by Edwards and Lauritzen are good references

Software

- as part of the Bioconductor project we are producing software for describing, rendering and interacting with graphs
- three R packages released
- graph: basic definitions/classes etc
- Rgraphviz: interface to graphviz
- RBGL: interface to the Boost graph library

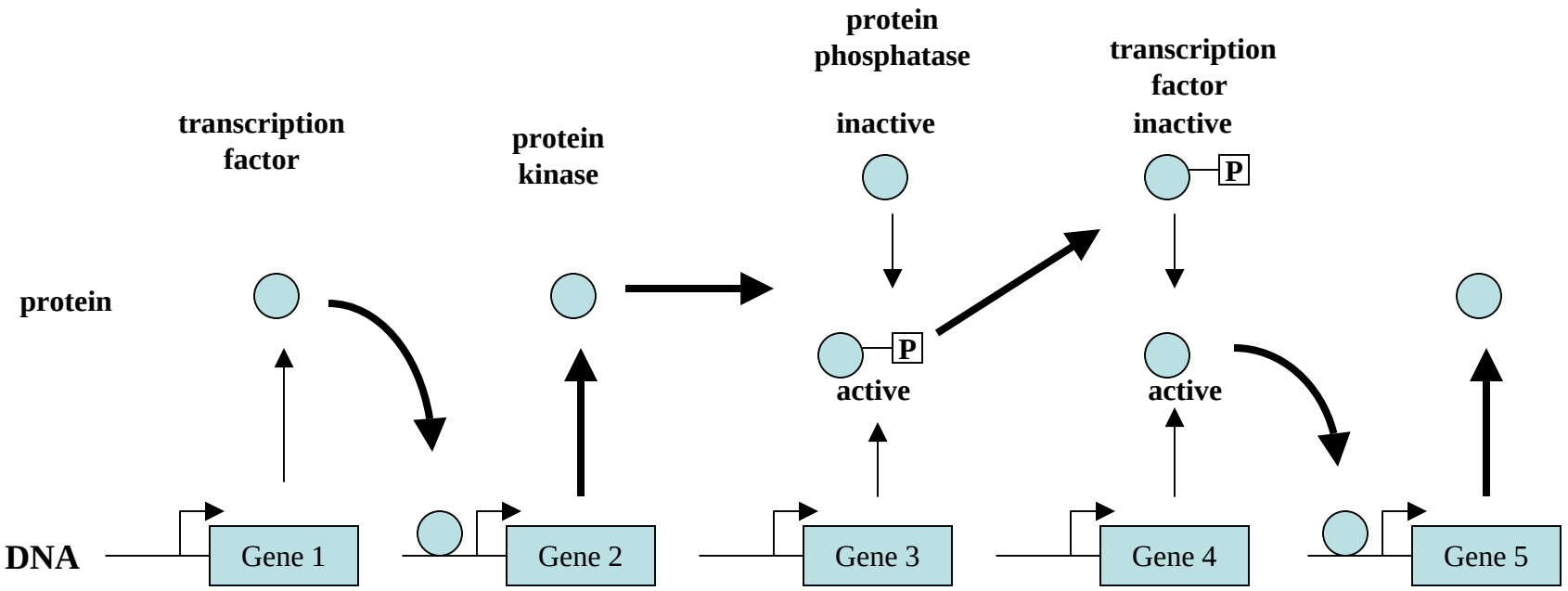
The Central Dogma

- DNA makes RNA (transcription)
- RNA makes protein (translation)
- the physical operations and interactions that are involved in these processes are very complex
- they almost always represent many to many relationships

Some Examples

- a transcription factor is a gene product that enhances or inhibits the transcription of other genes
- transcription factors are not generally specific (they have many targets)
- these targets have many targets …

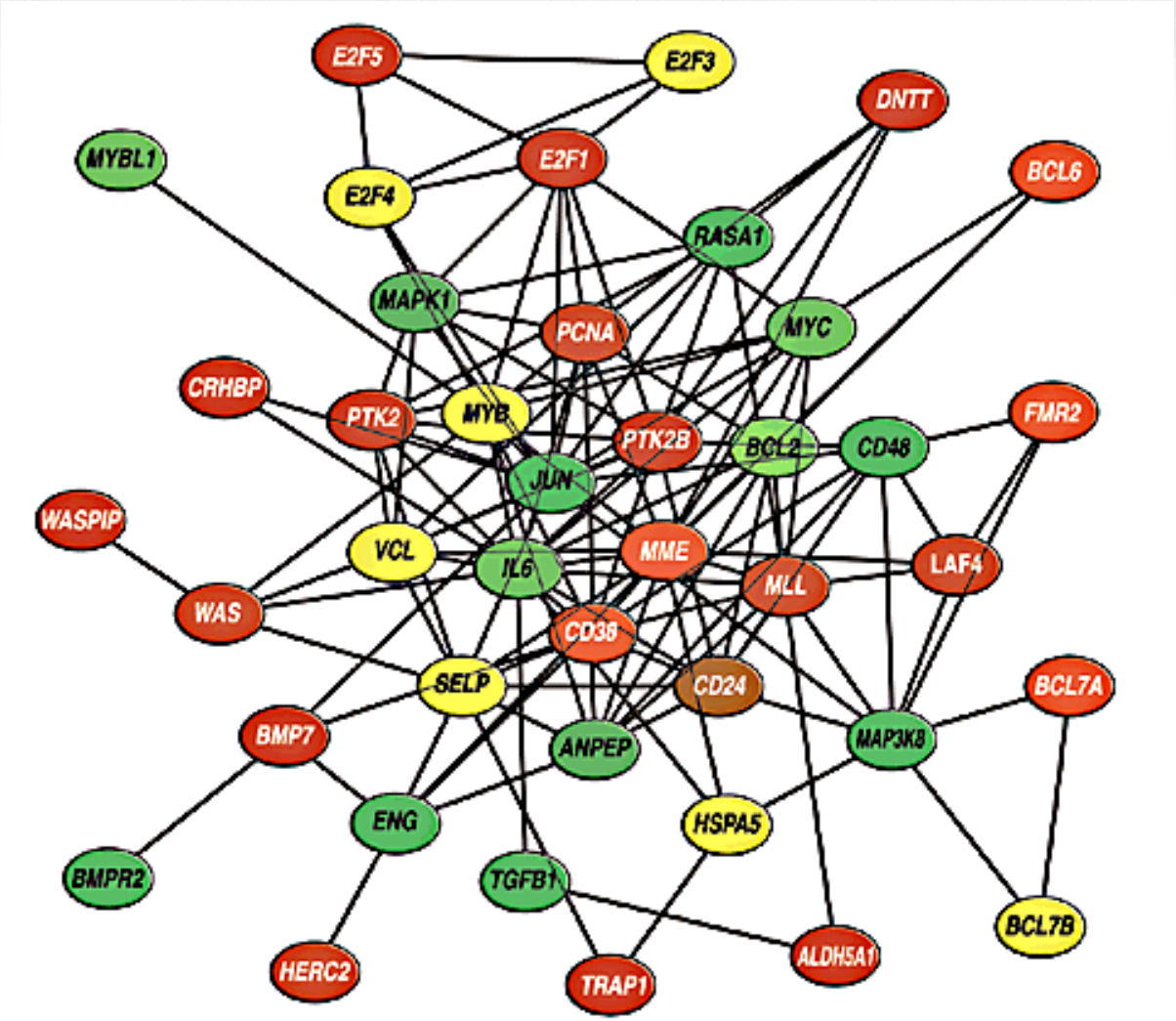
An example of the interactions between some genes (adapted from Wagner 2001)



Medical Literature

- co-citation in papers often indicates a relationship
- a paper may discuss multiple genes; each gene may be documented in multiple papers
- what graph are we interested in?
- what graph do we have data about?

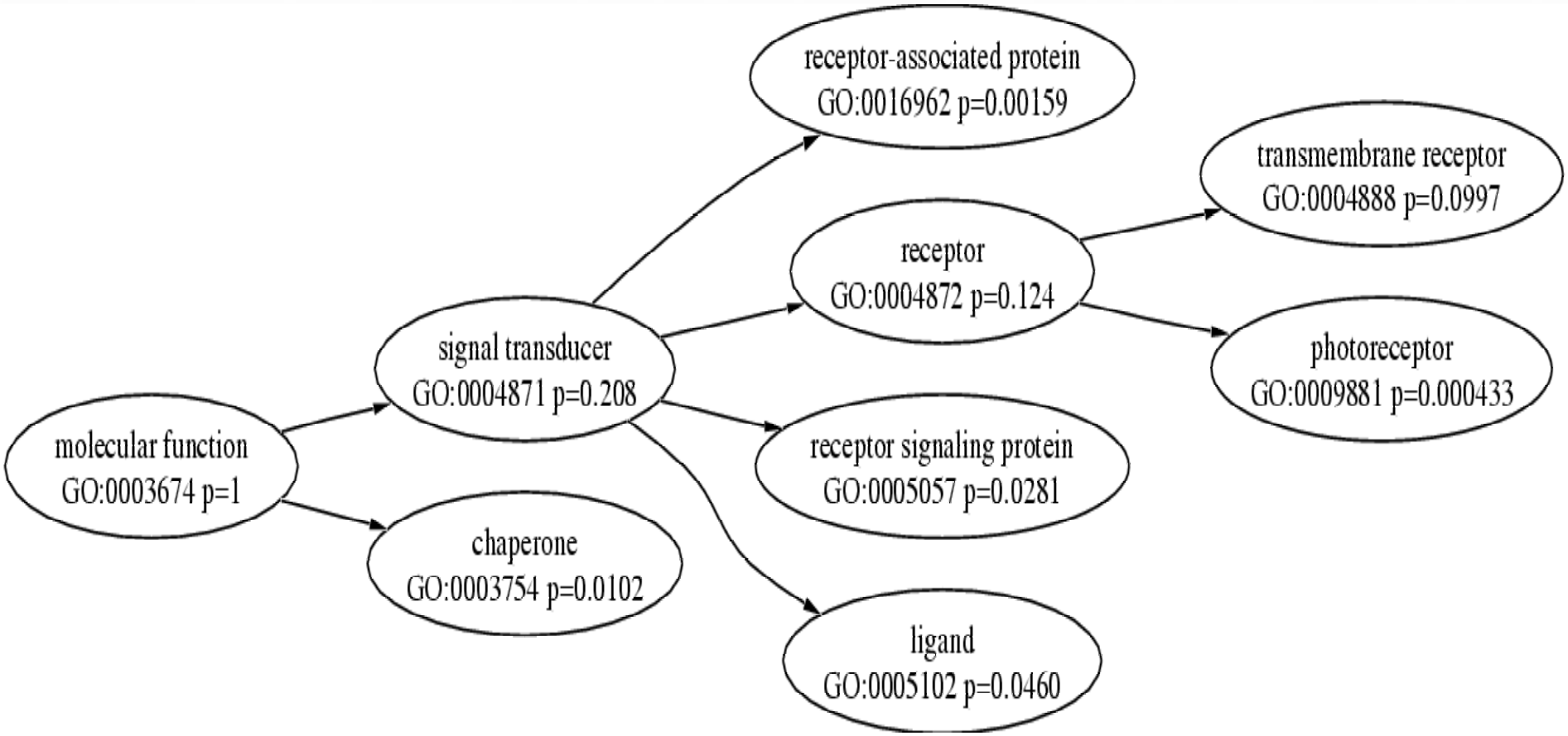
From Masys et al.



Gene Ontology

- Gene Ontology Consortium: a set of terms (or vocabulary) for documenting molecular function, cellular component or biological process
- some method (an oracle) associates genes with terms
- a gene can be associated with multiple terms, a term has multiple genes (it is a bit more complicated)

Adapted from Lord *et al*



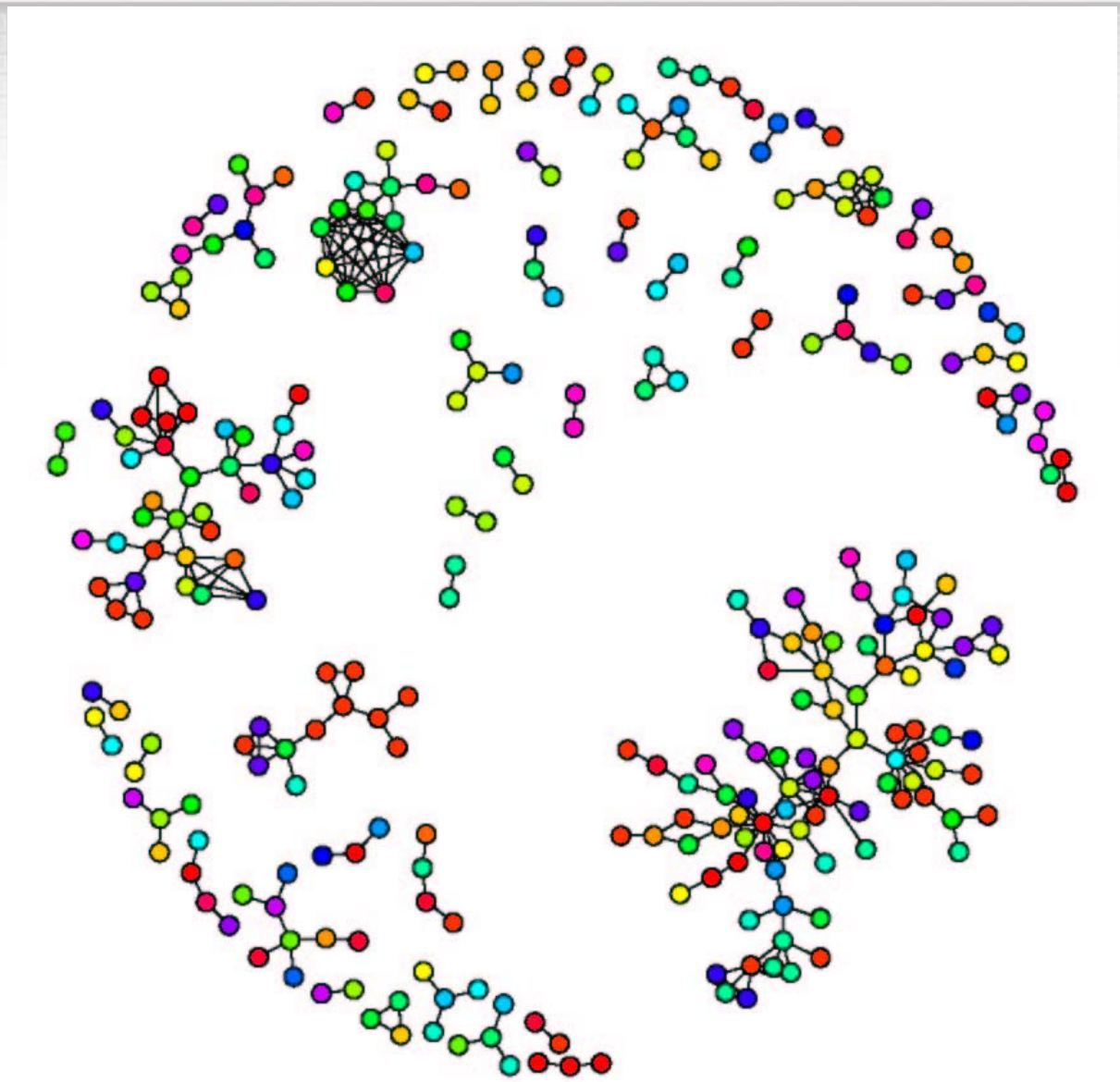
Protein-protein Interaction

- proteins seldom act individually
- they tend to act in pairs or groups to carry out their objectives
- some proteins are involved in many different groupings, others in only one
- different data sources (literature, MIPS, Y2H and TAP)

PPI

- are we interested in protein-protein interactions?
- are we interested in protein complexes that carry out biological processes?
- the data are usually consistent with the first question
- the inference is often oriented towards the second!

BIOCONDUCTOR
open source software for bioinformatics



Other Data Sources

- there are many other data sources available to us
- DNA microarrays, arrayCGH, SAGE, protein data, …
- how do we integrate these different data sources to better understand and explore the data at hand
- to focus the set of reasonable hypotheses and determine the next experiment

Combining Data

- there is a lot of evidence that there is an association between coordinated gene expression and participation in a protein complex
- in the last part of this talk we will directly address that question (raised in Ge *et al*, Correlation Between Transcriptome and Interactome…)

Basics

- a graph is a collection of vertices (V) and edges (E) between the vertices
- G=(V,E) to denotes the graph G
- |V| denotes the cardinality of the set V
- two vertices, v_i and v_j are said to be adjacent if they have an edge between them

Exploratory Data Analysis

- idea is to reveal structure or patterns in the data
- this depends on what you are looking for
- in classical statistics much of EDA is carried out with visualization methods
- with graphs/networks it is not yet clear what strategies will be useful

EDA

- graph layout is a hard problem
- it is often controlled by some form of specific optimization
  - minimum edge crossings
  - minimum edge length
  - etc
- but seldom optimized for information visualization

EDA

- there is a need for experiments, along the lines of those carried out by Cleveland and associates in the 1970s for visual perception
- what are you trying to show, does the audience see that?
- H. Purchase (UK) has done some experiments but more are needed

EDA

- does a graph conform or not to some sort of model?
- from a statistical or applications perspective graphs are being constructed on data – and are hence imperfect
- we must deal with missing edges:
  - edges that were not found
  - edges that were not looked for

Tools

- we can look at:
  - node characteristics
    - in and out degrees
    - notions of centrality
  - cohesive subgroups
    - cliques and near cliques
  - cut-points and cut-sets
    - separation

Tools

- the boundary of various subgraphs
- relationships to other graphs
  - intersection, union, complement
- often we are in the setting where we have multiple graphs all defined on the same set of nodes and so we have a different set of definitions for union, intersection, and complement than a mathematician might

Tools

- in addition to these static or structural properties there is clear benefit to interactivity
  - moving nodes/edges
  - collapsing node sets
  - interrogating nodes
  - interrogating edges
  - linked plots (brushing)

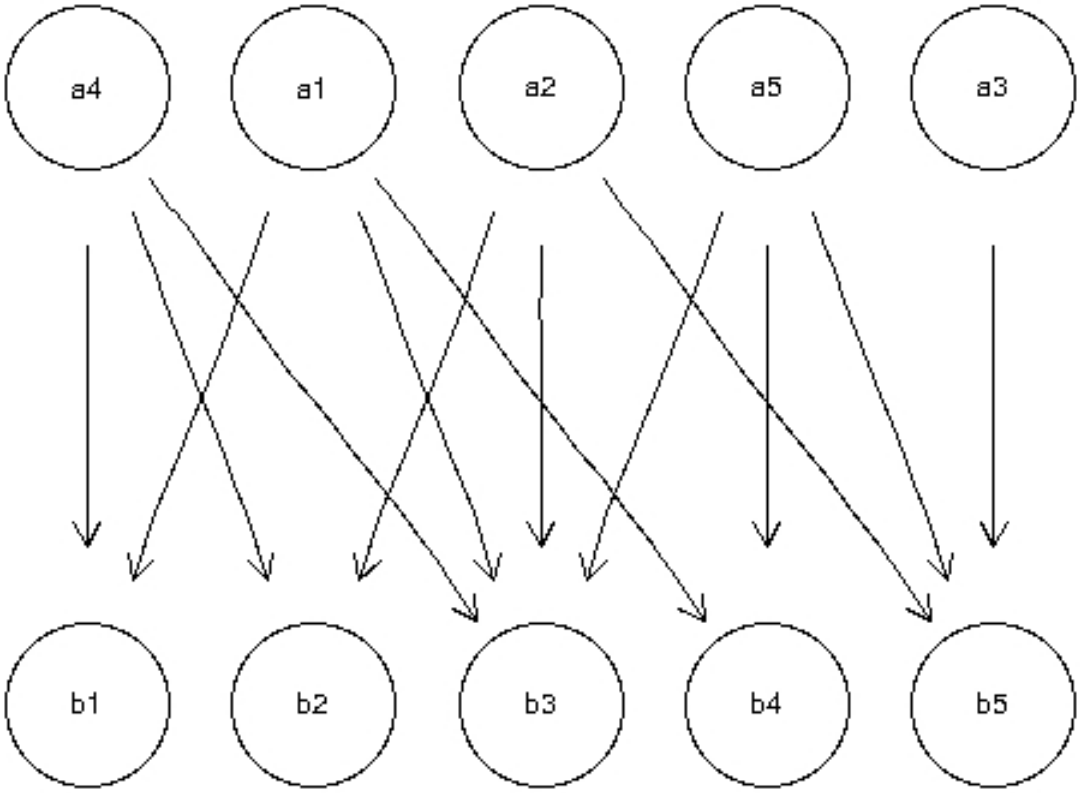
Bipartite Graphs

- if the nodes of a graph can be partitioned into two disjoint sets, $\boldsymbol{N}_1$ and $\boldsymbol{N}_2$, say
- such that all edges are between an element of $\boldsymbol{N}_1$ and an element of $\boldsymbol{N}_2$ (ie. all edges go from one set to the other; no within-set edges)
- then the graph is called a bipartite graph

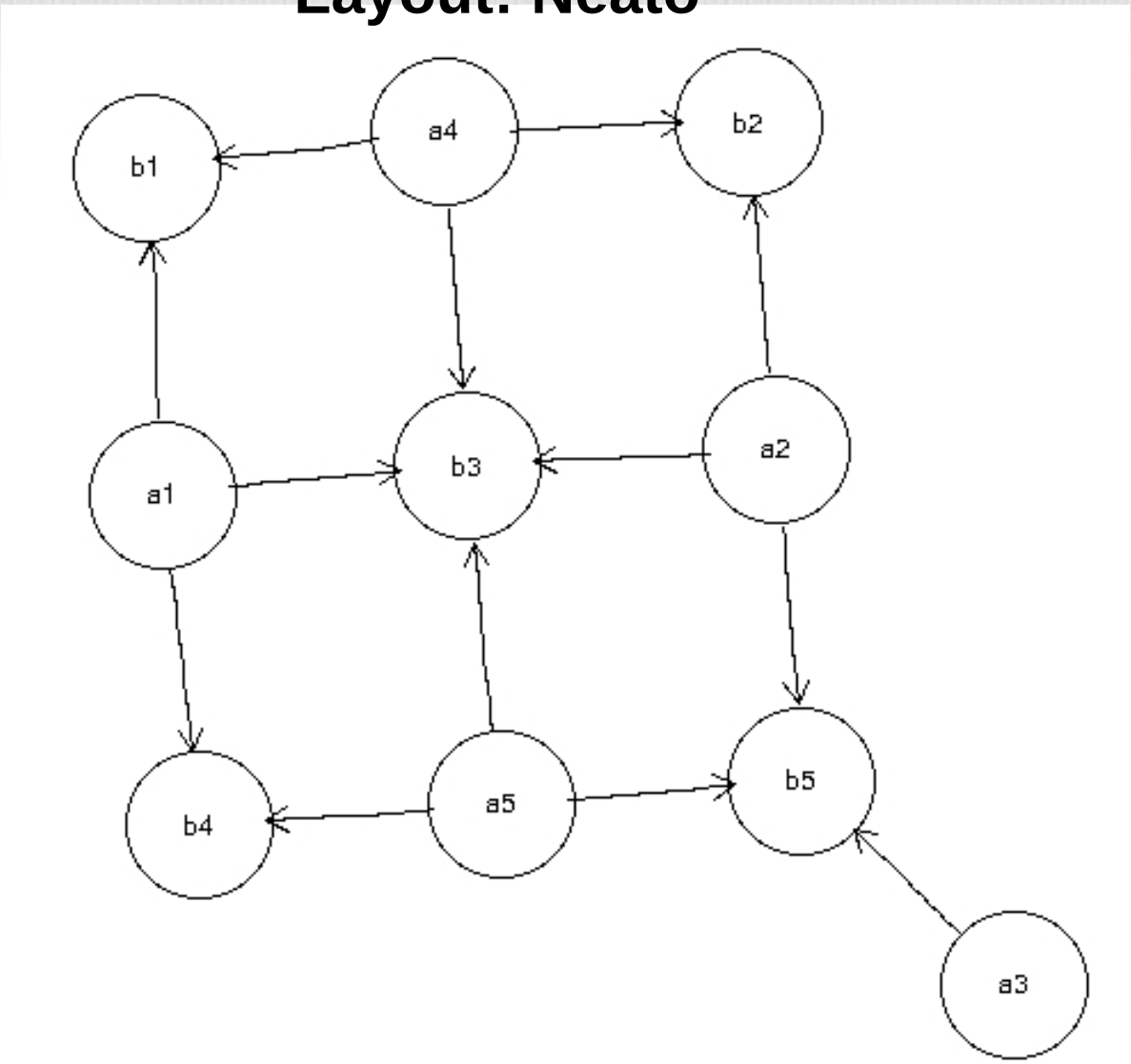
Layout: Dot

Genes

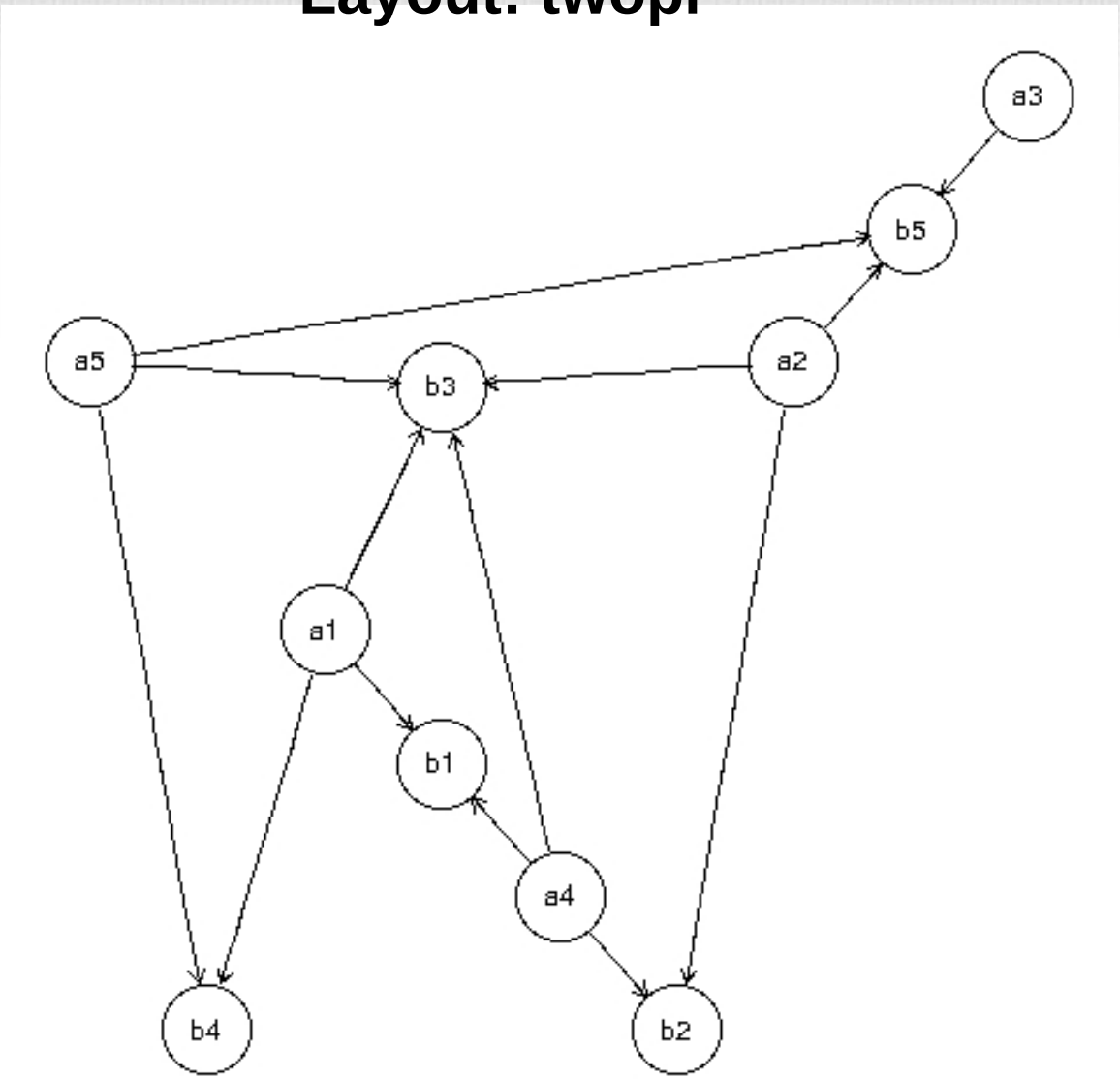
Papers



Layout: Neato



Layout: twopi



Bipartite Graphs

- how should we layout bipartite graphs?
- horizontal? vertical?
- minimize edge crossings?
- order from left to right according to degree for the top and then for the bottom either to minimize crossings or by degree?
- what are we trying to see in this?

Affiliation Networks

- in social network analysis a bipartite graph that associates individuals (actors) with events is often called an affiliation network
- we will use the term single-mode graph when we are interested in understanding properties about one type of node (either actors or events)

Affiliation Networks

- two examples of biological affiliation networks:
  - genes are one type of node and papers that discuss those genes are the other
  - genes/proteins are one type of node and protein complexes are the other

Affiliation Networks

- the adjacency matrix for an affiliation network is N by M (where N is the number of nodes of the first type and M the number of nodes of the second)
- the matrix is filled with zeros and ones
  - a one in row i column j indicates that individual i participates in activity j
- we will label this matrix ***A***

Affiliation Networks

- interest often focuses on either the rows (genes) or the columns (papers/protein complexes)
- a one-mode graph is obtained by considering the matrix product ***AA'*** or ***A'A***
- in many cases the matrix multiplication is Boolean (we only see 1's and 0's in the matrix products)
- the diagonal is often not interesting (observed)

Affiliation Networks: PubMed

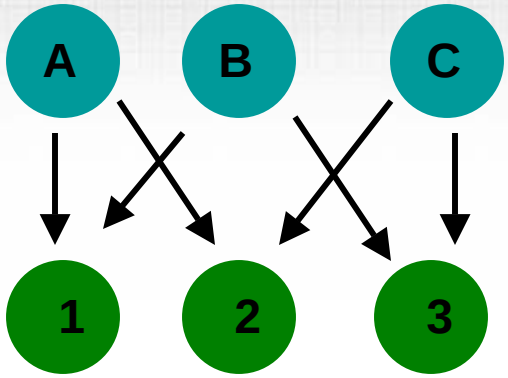
- we can derive a graph on genes where edges are created between genes that share citations
- or a graph on papers where the edges represent shared genes
- in both cases the resulting graph is undirected and valued

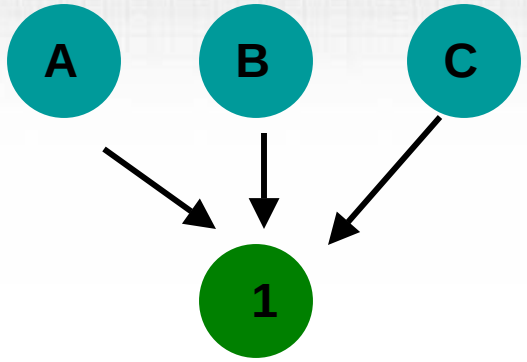
Affiliation Networks

- edge weights could be important
- in the gene/paper graph we might want to down-weight papers that have lots of genes
- we might think of each paper as having constant weight/impact and so if paper j has in-degree m then each in-edge receives weight 1/m

Affiliation Networks

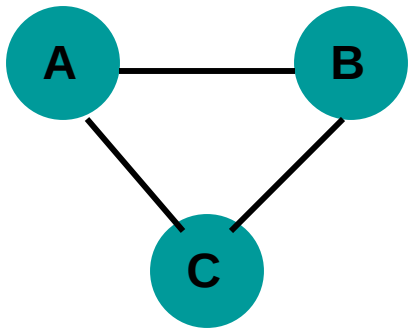
- because the one-mode graphs are constructed by using pairwise information (shared nodes of the other mode) you can only make pairwise inference from them
- thus, cliques and other subgroups in the one-mode graphs can arise in many (undetectablely) different ways


A
B
C
1
2
3


A
B
C
1

Both give the same one-mode (blue) graph

Both give the same one-mode (blue) graph



Affiliation Networks

- if we see a clique in the PubMed graph between genes A, B and C we cannot tell from that source alone whether there were three papers that cited pairs or one paper that cited all three
- if all we see is the single-mode graph our inference must be restricted to pairwise relationships

Affiliation Networks

- let us consider protein-protein interactions
- the general objective is to identify protein complexes
- that is, sets of two or more proteins that form a unit that carries out a particular biological objective
- a number of technologies are appearing that provide data of this sort

Tandem Affinity Purification

- TAP data arise from a bait-prey experiment (Gavin *et al*, Ho *et al*)
- marked proteins are used as *bait,* they are introduced into the cell and then retrieved together with all things that they interacted with
- in a sense, the observed data are of the form of ***AA'*** and we want to know about ***A***

TAP

- but the map from ***AA'*** to ***A*** is one-to-many so some statistics are needed
- more importantly the relationships are not quite so simple
- there are three types of edges
  - edges found
  - edges not found and probed for
  - edges not found and not probed for

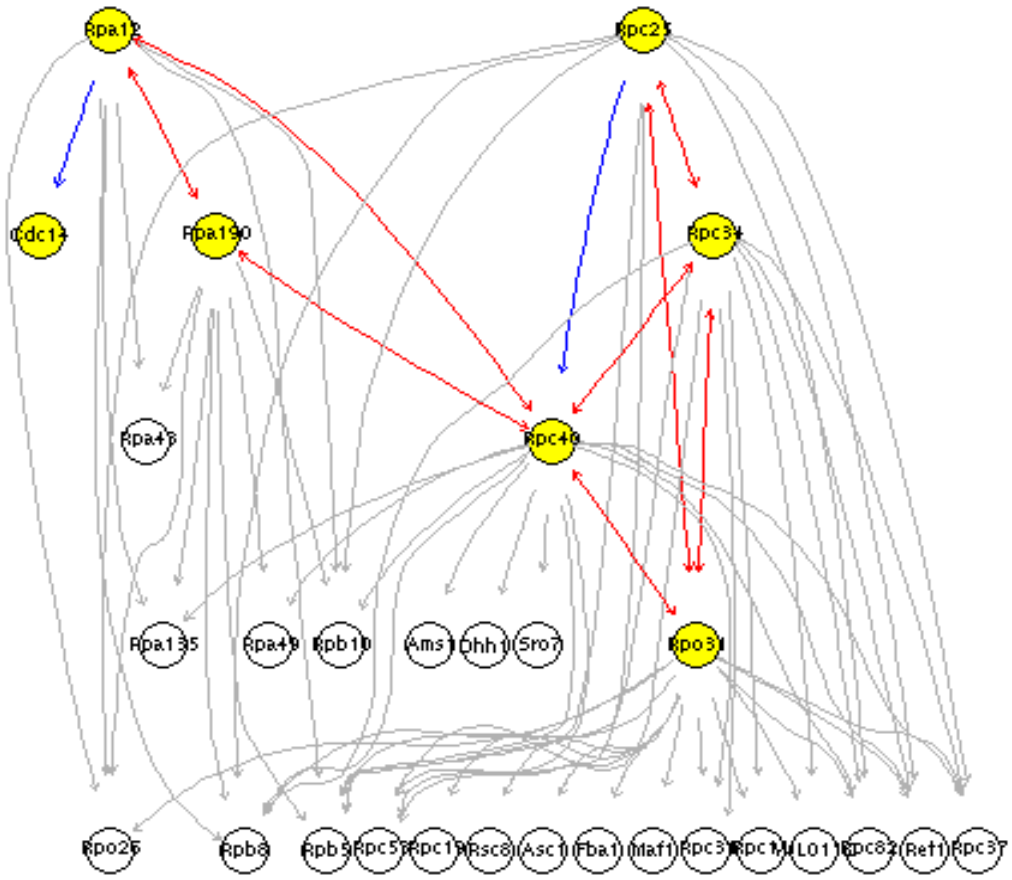
TAP

- a protein used as a bait has looked for all other proteins (and hence all edges)
- some experimental error is involved (as well as some structural issues) so that the resulting edges are imperfect (found but not real and real but not found)
- proteins not used as baits can only have in-edges

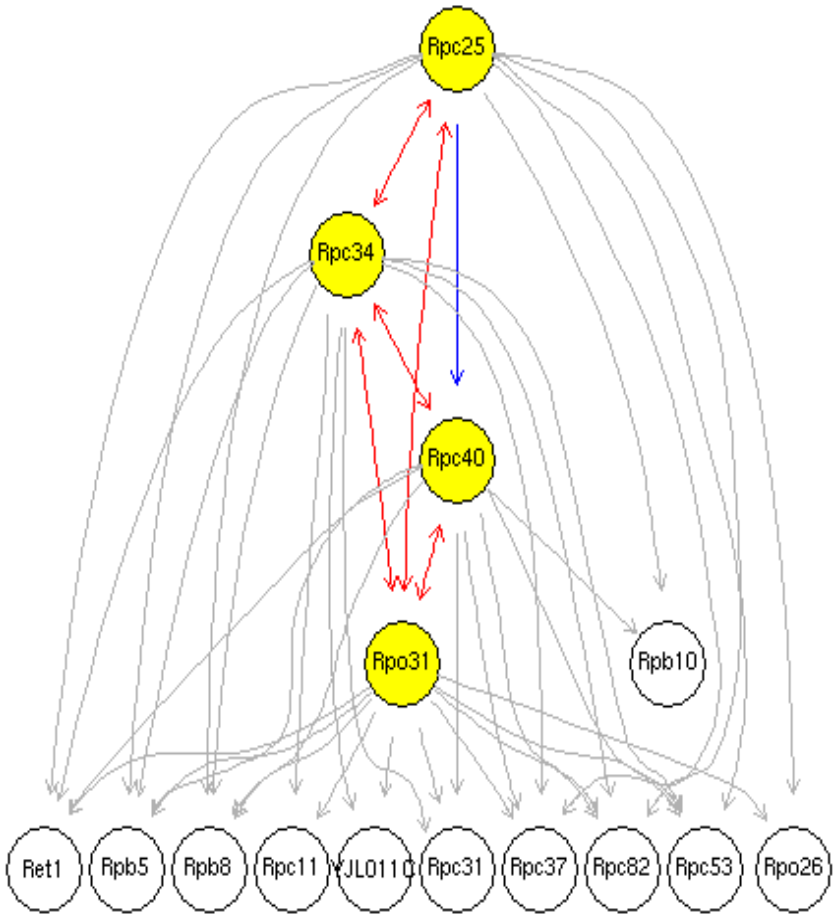
TAP

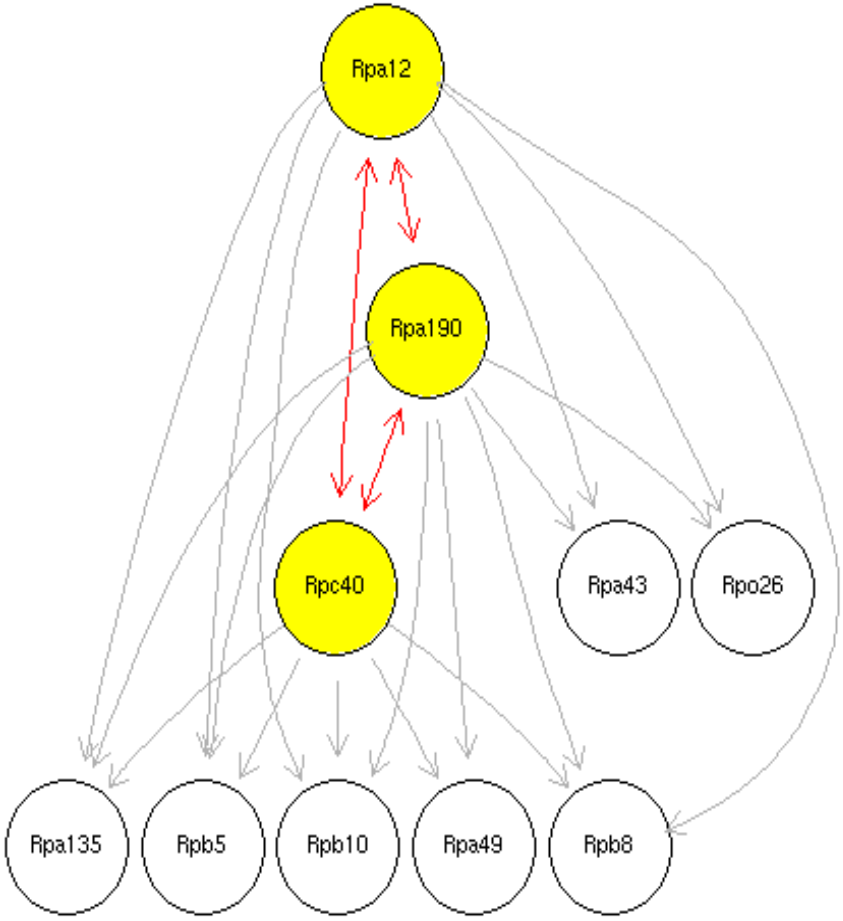
- the next few pictures represent a protein complex
- red edges represent reciprocity
- blue edges indicate that one found the other (bait to bait)
- gray edges represent bait to prey relationships

Gavin *et al*


Rpa12
Rpc25
Cdc14
Rpa190
Rpc34
Rpa43
Rpc40
Rpa135
Rpa49
Rpb10
Ams1
Dhh1
Sro7
Rpo31
Rpo26
Rpb8
Rpb5
Rpc5
Rpc11
Rsc8
Asc1
Fba1
Maf1
Rpc31
Rpc19
LO11
Rpc82
Ret1
Rpc37

Scholtens


Rpc25
Rpc34
Rpc40
Rpo31
Rpb10
Ret1
Rpb5
Rpb8
Rpc11
JL011C
Rpc31
Rpc37
Rpc82
Rpc53
Rpo26


Rpa12
Rpa190
Rpc40
Rpa43
Rpo26
Rpa135
Rpb5
Rpb10
Rpa49
Rpb8

Complexes

- So, what is a complex?
- the first picture is representative of what you will get from MIPS (and other sources) and is based on data from Gavin *et al*.
- the second is due to work with D. Scholtens, and there are 4 papers that support the existence of 3 complexes based on these proteins (one not in the data)

TAP

- by making better use of the data (different types of edges) we identified two clusters rather than one
- we also use data on cellular location of the proteins in our model
- this observation (two not one) is supported by the literature

Ge *et al* – PPI and Transcriptome

- they asked an interesting question
  - is there a relationship between gene expression (from a time course experiment) and which proteins interact?
- data from a microarray experiment were clustered
- two PPI data sets (literature and y2h) were used to ask whether there are more within group PPI than between group PPI

Interactome-Transcriptome

- this can be phrased as a question about graphs
- the clusters can form a graph
  - all genes in the same cluster have edges
  - there are no edges between clusters
- now we can easily identify within and between cluster interactions by standard operations on graphs

Interactome-Transcriptome

- the intersection of the cluster-graph and the PPI graph yields within cluster edges
- we can take the clusters, find the induced subgraphs and attribute edges per cluster

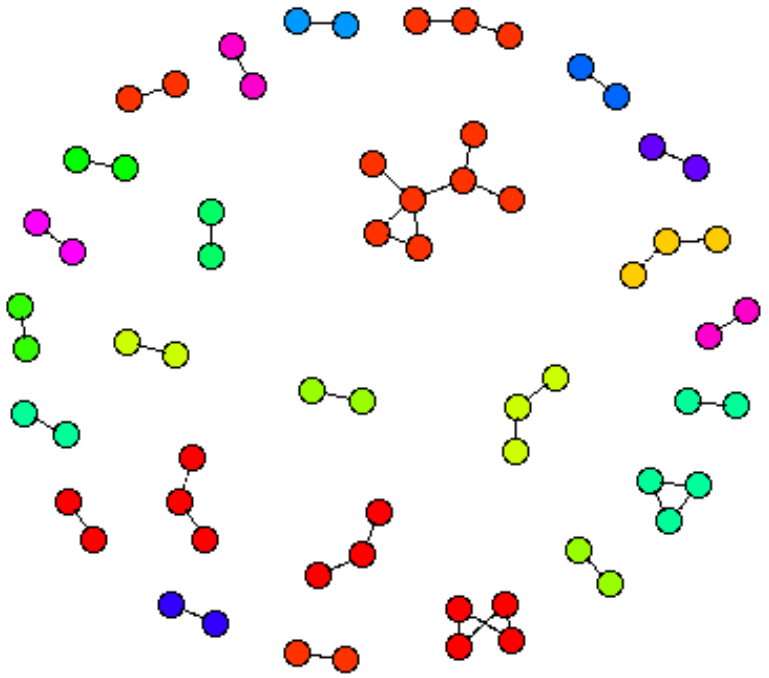
The Literature Cluster interactions

Some obvious questions:

which clusters have lots?

which have few?

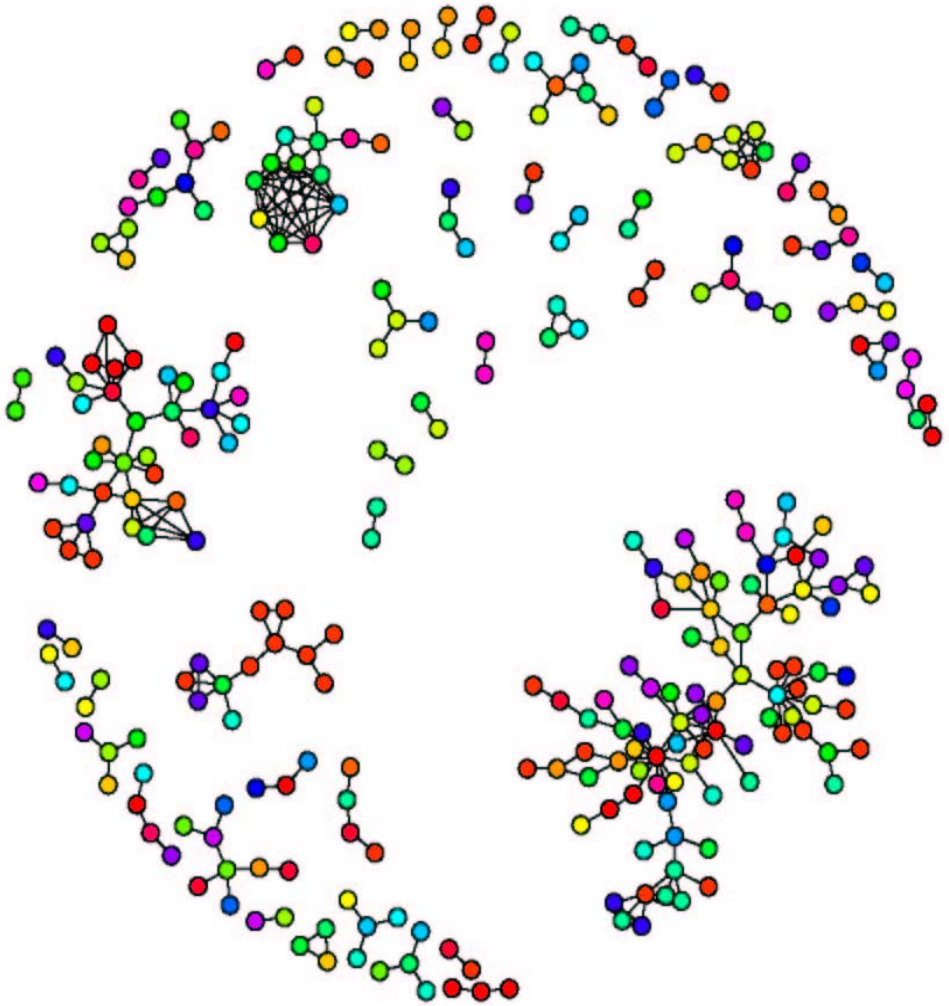
are there other edges and where?



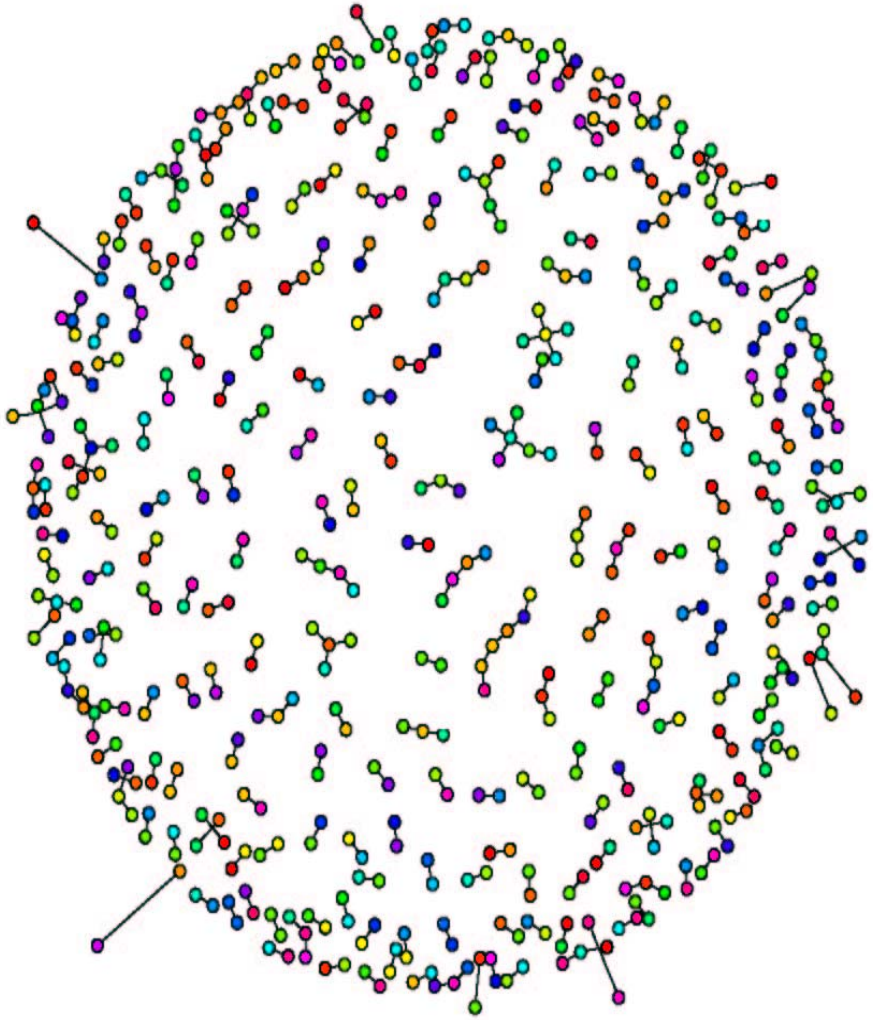
Computational Biology

- to test the hypothesis that there was a relationship between the transcriptome and the interactome they tested the hypothesis that there were more edges within clusters than you would expect by chance.
- their test was based on the Erdos-Renyi model for random graphs
- random edge model

Observed Data



A realization from the Erdos-Renyi model



A different model

- it might be better to keep the subgraph structure and permute the node labels
- this is basically a conditioning argument
- with the permuted node labels compare to the clusters (fixed) and count the number of within cluster edges
- note the symmetry with permuting the labels for the clusters and keeping the graph fixed

Inference

- a test the independence of the row classification and the column classification can be phrased in terms of graphs G1 and G2
- we can apply either the hypergeometric test (Erdos-Renyi model) or the node label permutation test
- in some examples the node-permutation method is equivalent to Fisher's exact test!

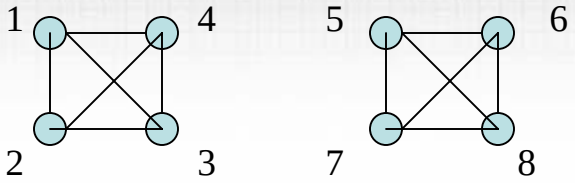
	C	D	
A	3	1	4
B	1	3	4
	4	4	8

we can view this as data on 8 individuals

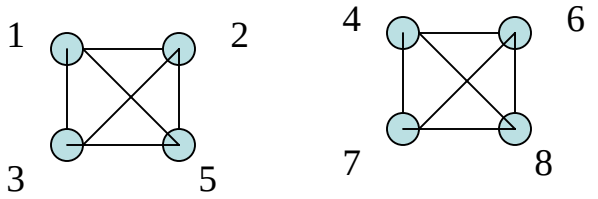
the row and column totals should be conditioned on

there are 8 nodes and 28 edges in the complete graph

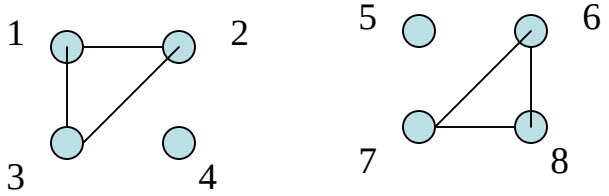
Row Graphs



Column Graphs



Intersection



Inference

- the row graph has 12 edges, as does the column graph
- for the Hypergeometric distribution we have (28, 12, 12) as parameters
- but this ignores the structure – the row (or column graphs) have 12 edges by virtue of being two clusters of size 4
- the edges are not random

Inference

- the random permutation of node labels (in either graph) yields Fisher's exact test
- it would be nice to explore the other connections that arise from considering the commonalities between the graph approach and standard independence testing

Conclusions

- describing the questions (and data) in terms of graphs greatly simplifies the analysis – in the sense that I just think about operations on graphs
- graphs present many computational, analytic and graphical challenges (opportunities)

Thanks

- Jeff Gentry
- Vincent Carey
- Wolfgang Huber
- Emden Ganzner
- Stephen North
- Denise Scholtens
- Beiying Ding
- Elizabeth Whalen
- Duncan Temple Lang
- Jianhua Zhang