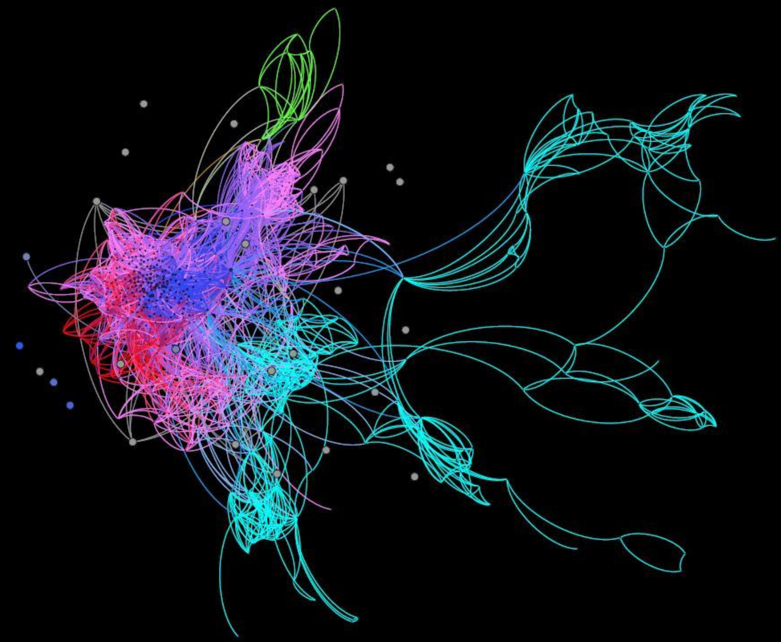
On Classifying **Complex Networks** by their **Topological Features**



Marina von Steinkirch

May, 6th 2014

github.com/mariwahl/MLNet-Classifying-Complex-Networks
github.com/mariwahl/MNet-Network-Analysis

What Complex Networks ?

Databases:

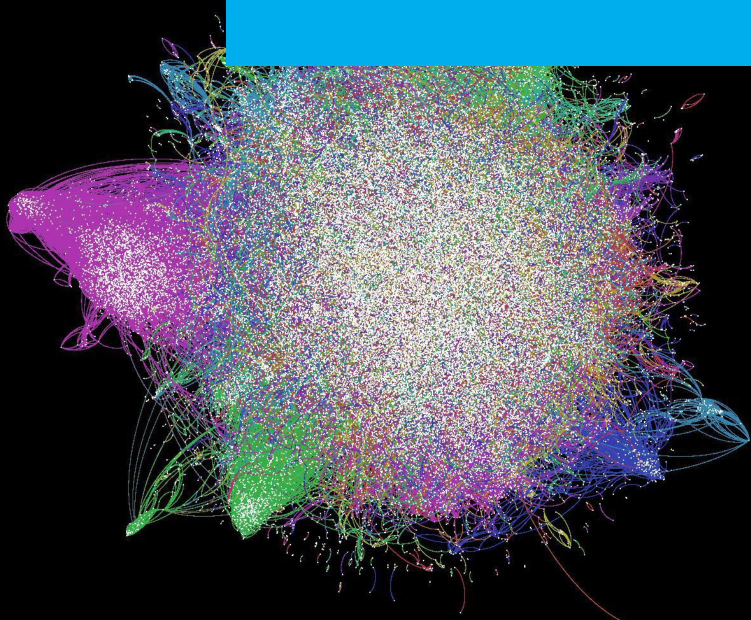
- KONECT Database, http://konect.uni-koblenz.de/networks
- SNAP Database, http://snap.stanford.edu/data
- VLADO database, http://vlado.fmf.uni-lj.si/pub/networks/

divided in 4 groups...

Social Networks

Brightkite: Location-based social
networkin
users shar
checking-

(58,228 n



We collected:

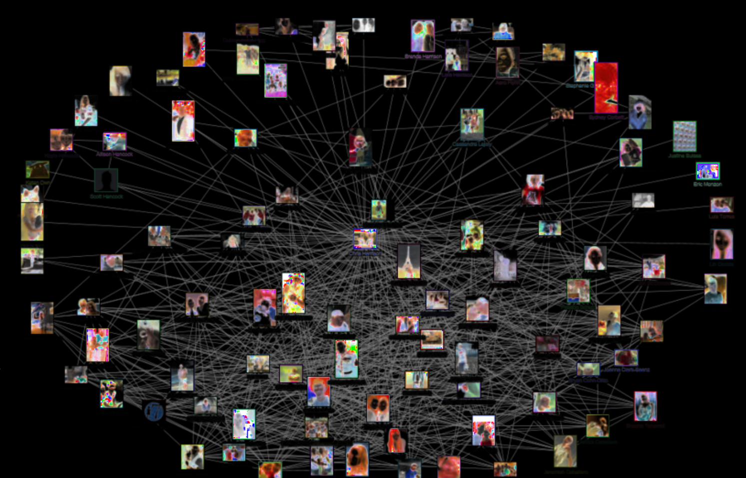
- 5 Social networks with geographic check-

ebook).

rust).

Total: 992

Facebook:
Friend lists.
(4039 nodes
and 88234
edges).



Biological Networks

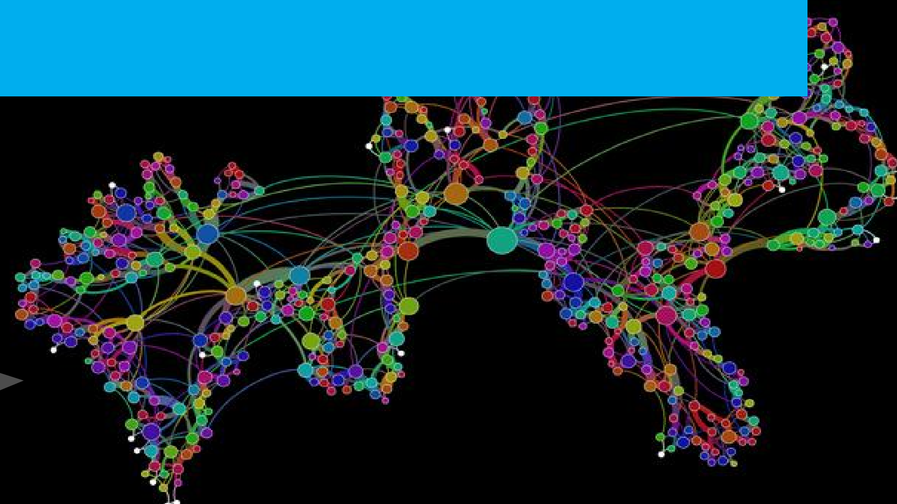
Yeast protein, http://www.simonsfoundation.org/

We collected:

- 2 Carbon exchanges.
- 43 Cellular (substract in cellular networks).
- 43 Metabolic networks (interactions betw
- 3 Ye
- 8 At

Total: 99

Edges are usually symmetrical and directed!



Metabolic processes: Hierarchical modularity of nested bow-ties in metabolic networks, Zhao et al 2010

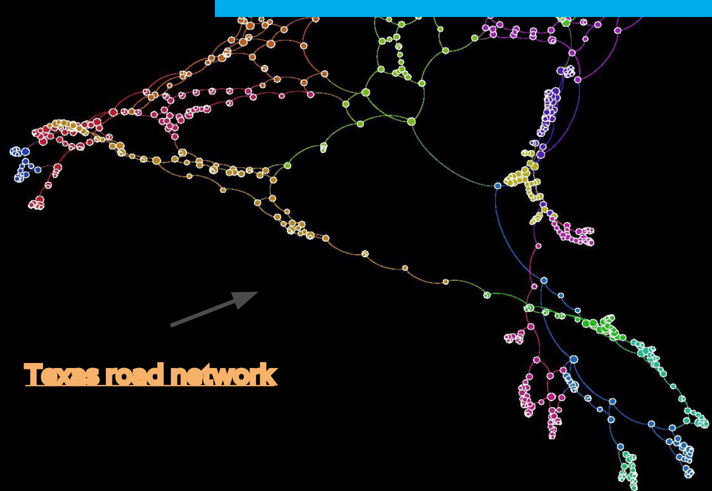
Technological Networks

Autonomous Systems (AS) peering information inferred from Oregon route.

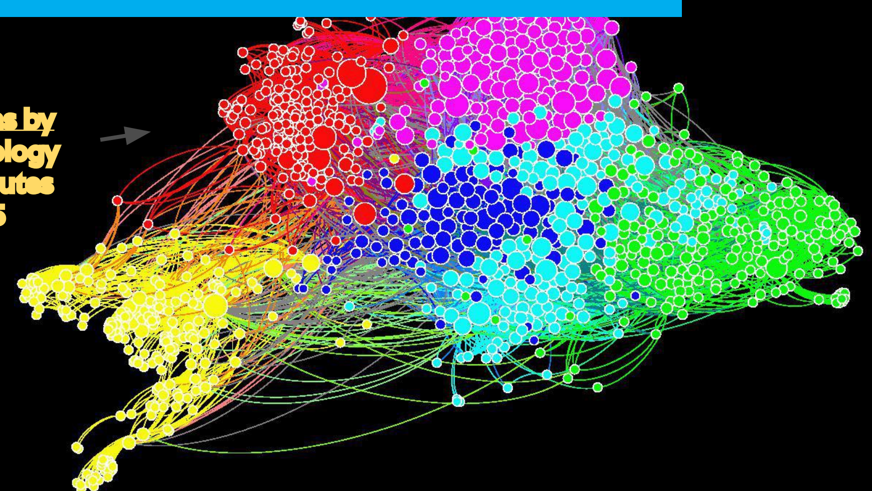
We collected:

- **117** Autonomous systems (graphs of the internet).
- **7** Roads (nodes represent inte
con

Total: 124



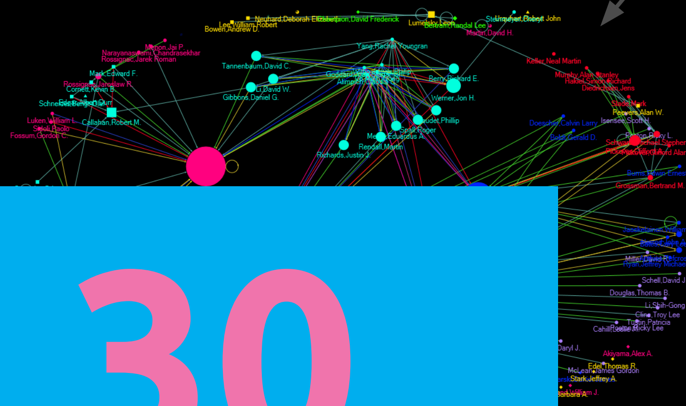
Texas road network



Autonomous systems by Skitter: Internet topology graph. From traceroutes run daily in 2005

Information Networks

US Patent Citation Data Set available at http://www.nber.org/patents

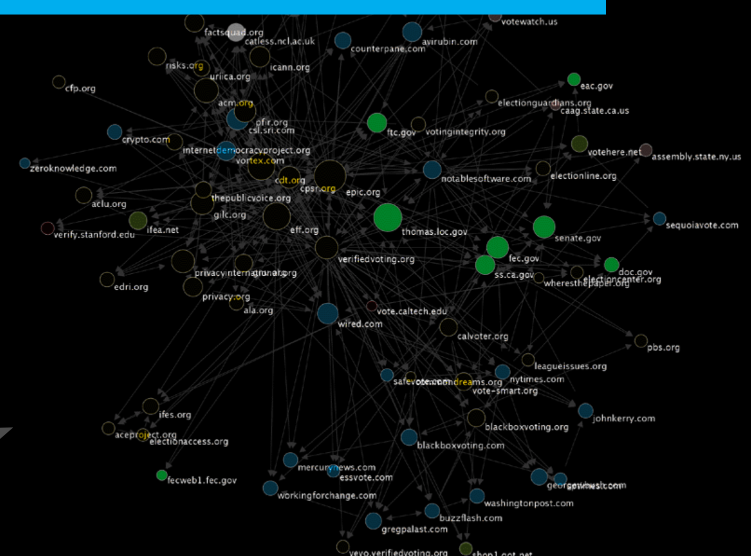


We collected:

- 2 Citation (nodes represent papers, edges represent citations).
- 8 Colla… scienti…
- 3 Con… netwo…
- 4 Web… edges …
- 4 Amazon Product Review.
- 9 Peer-to-peer.

Total: 30

Knowledge based networks: data linked together!



http://farrell.org/webgraph/research/evote.html

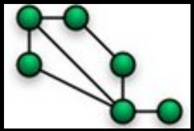
Total number of Complex Networks

1245

github.com/mariwahl/MNet-Network-Analysis

Graph Topological Features?

Topological Features…



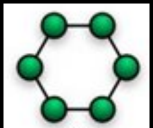
Density (den): The ratio of the number of edges and the number of possible edges.

Clique number(cqn): Return the size of the largest clique for G.

23!

(global and local/average)

Communicability centrality (pr): For a *node* n, it is the sum of closed walks of all lengths starting and ending at node n.



Edge connectivity (eco): Minimum number of edges that must be removed to disconnect G.

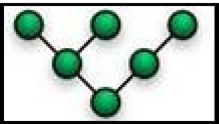
Density (den): Ratio of existing to possible links in G. It ranges from no link at all to all nodes connected (0 and 1 respectively): $den(G) = \frac{m}{n(n-1)}$. Real networks are usually very sparse, with ~ 0.1.

Coreness (cor): A k-core is a maximal subgraph that contains nodes of degree k or more. The *cor* of a node is the largest value k of a k-core with that node.

Expansion (ex): An expander graph is a sparse graph that has strong connectivity properties, so that the complete graph has the best expansion property. A graph is a good expander if it has low degree and high expansion parameters.

Closeness Centrality (cc): Measures how fast information spreads from a given node to other reachable nodes in the graphs. For a *node* u, it represents the reciprocal of the average shortest path length between u and every other reachable node in the graph: $cc(u) = \frac{n-1}{\sum_{v \in \{V_u\}} d(u,v)}$, where $d(u,v)$ is the length of the shortest path between the nodes u and v.

Eccentricity (ecc): Represents, for a *node* u, the maximum length of the shortest path between u and every other node in G: $ecc(u) = max_{v \in V} d(u,v)$. If u is isolated, then $ecc(u) = 0$.



Degree (deg): For a *node*, it is defined as the number of its neighboring edges. It can be formally defined using the *adjacency matrix*: $deg(u) = \sum_{v \in V} a_{uv}$. In real-world networks, the average degree often follows a power law (*scale-free* networks).

Pagerank (pr): Computes a ranking of the nodes in the graph G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages.

Minimum Effective Eccentricity or Radius (rad): Represents the minimum value of *ecc* over all nodes in the graph G: $rad(G) = min\{ecc(u) | u \in V\}$.

Betweenness Centrality (bc): For a *node* u, it is the sum of the fraction of all-pairs shortest paths that pass through u. If we denote σ_{vw} as the total number of shortest paths between v and w, and $\sigma_{vw}(u)$, the total number of shortest paths between nodes v and w going though u, the betweenness centrality is $bc(u) = \sum_{v<w \neq w} \frac{\sigma_{vw}(u)}{\sigma_{vw}}$.

Number of Edges (m): The total number of edges in the network (graph size), $m = |E|$.

Node connectivity (nco): Minimum number of nodes that must be removed to disconnect G.

Assortativity (r): Measures the similarity of connections in G with respect to the node degree. Graphs that have only single edges between vertices tend (in the absence of other biases) to show disassortative mixing by degree because the number of edges that can fall between high-degree vertex pairs is limited. Since most networks are represented as simple graphs this implies that most should be disassortative

Dispersion (di): Represents a variation from the mean values and identifies patterns.

Number of cliques (ncq): A clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge.

Percentage of Isolated Points (isop): The ratio of isolated points to the total number of nodes. An isolated point in G is a node with a degree zero.

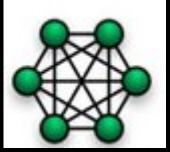
Maximum Effective Eccentricity or Diameter (diam): Represents the maximum value of *ecc* over al nodes in the graph G: $diam(G) = max\{ecc(u) | u \in V\}$.

Square clustering coefficient (scc): While *clc* gives the likelihood that any two neighbors of u are connected, *scc* gives the probability that two neighbors of node v share a common neighbor different from v.

Clustering coefficient (clc): For a *node* u, represents the likelihood that any two neighbors of u are connected: $clc(u) = \frac{2e}{k_u(k_u-1)}$, where k_u is the number of neighbors of u and e_u is the number of connected pairs of neighbors. If all the neighbors nodes of u are connected, then, the neighborhood of u is complete and $clc = 1$. If no nodes in the neighborhood of u are connected, $clc = 0$.

Transitivity (tra): A global measure of *clc*, it computes the fraction of all possible triangles present in G. The transitivity ranges from 0.1 to 0.8 in the real world network. It can be interpreted when picking a randomly node, as the probability for two of its neighbors to be connected.

Let us note $\gamma(G)$ the number of subgraph with 3 links and 3 nodes and $\tau(G)$ the number with at least 2 links and 3 nodes, then: $traG) = \frac{\gamma(G)}{\tau(G)}$.



But…

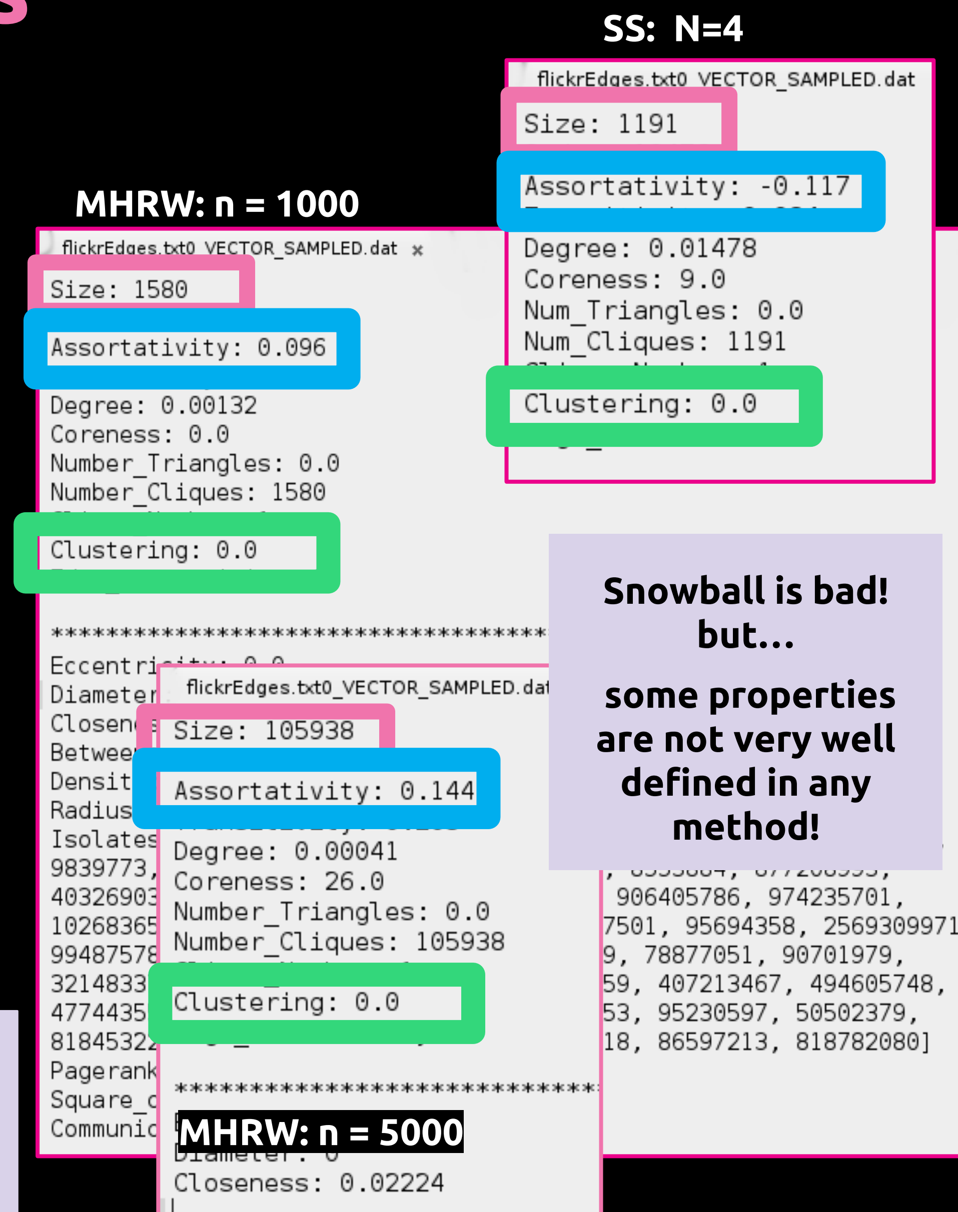
Features are
size-dependent!

Need to normalize
each graph!

Sampling Methods

★ Two types:
- Snowball Sampling (SS).
- Metropolis-Hasting Random Walk Sampling (MHRW).

★ Approaches:
- Graph orders: n = 500, 1000, 1500, 2000, 3000, 5000.
- Depth N =3, 4



Snowball is bad! but…

some properties are not very well defined in any method!

Example, the “online communication” network:

The entire network:
- Nodes: 106722
- Edges (size): 2316668
- Clustering: 0.001
- Assortativity: 0.144

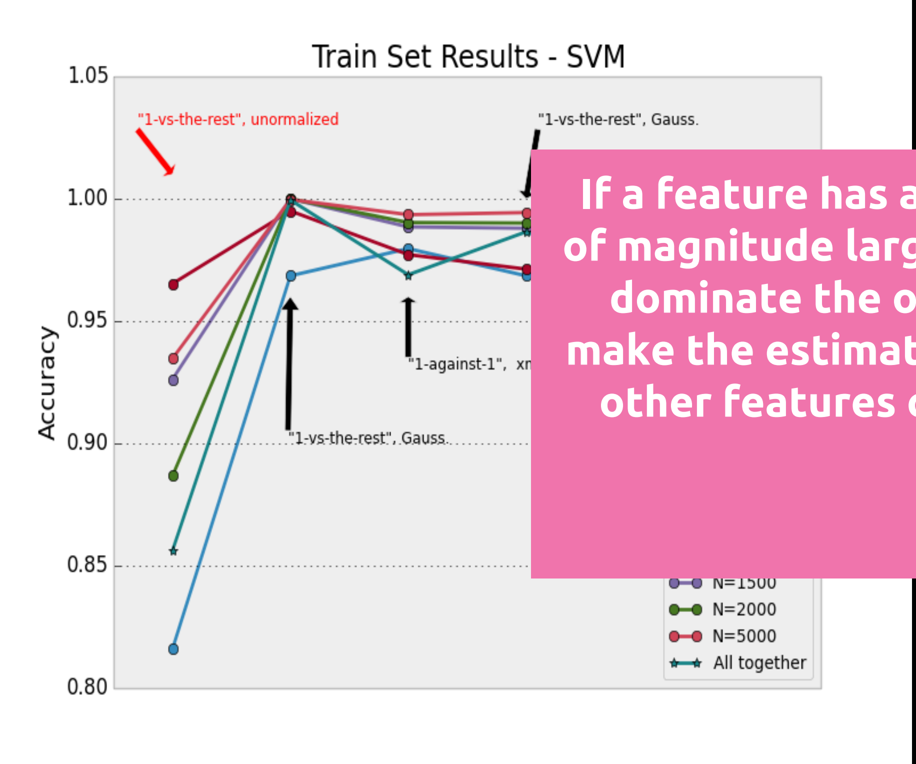
SMV it!

SVM approaches:

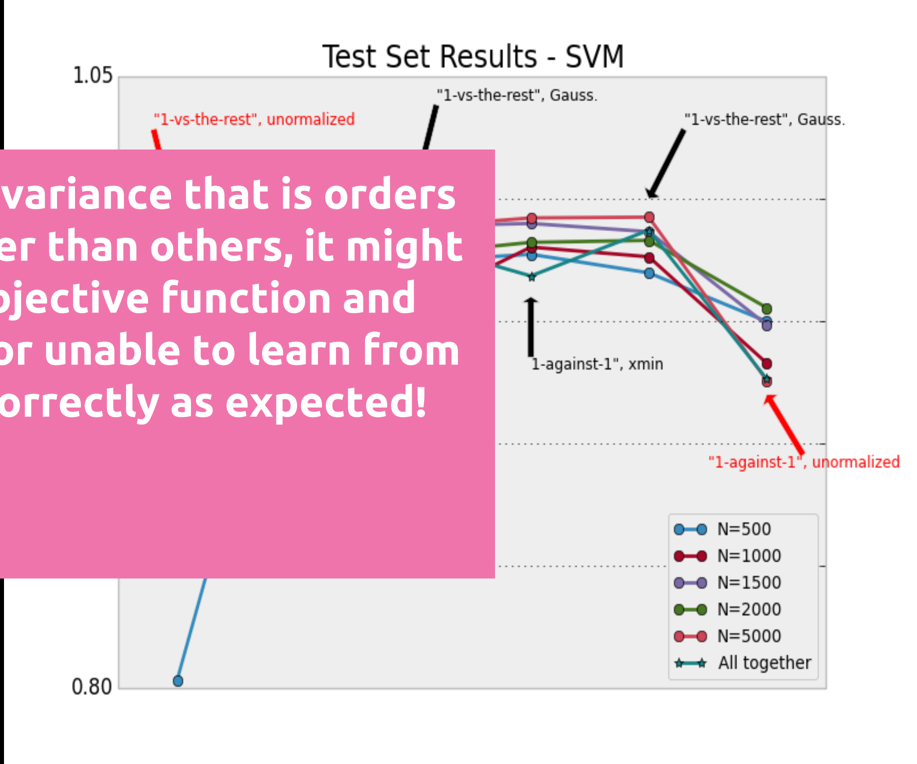
1. ‘one-vs-one’ (Knerr et al., 1990): n_class * (n_class - 1) / 2 classifiers are constructed and each one trains data from two class (kernel RBF)
2. ‘one-vs-all’: training n_class models (Linear)

Standardization/Scaling:

1. gaussian with zero mean and unit variance
2. minimum and maximum value

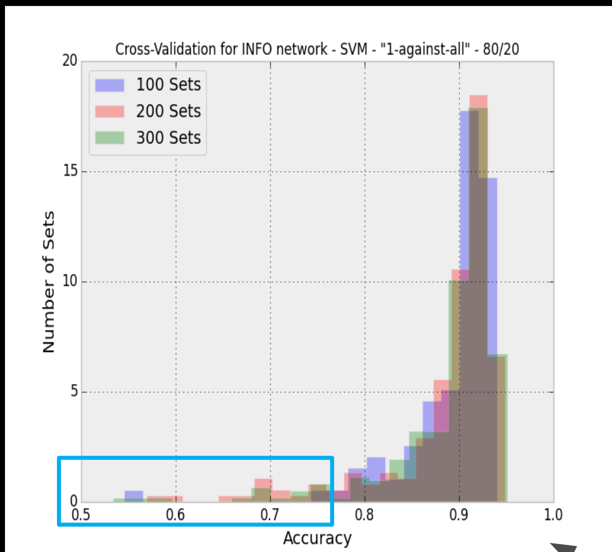


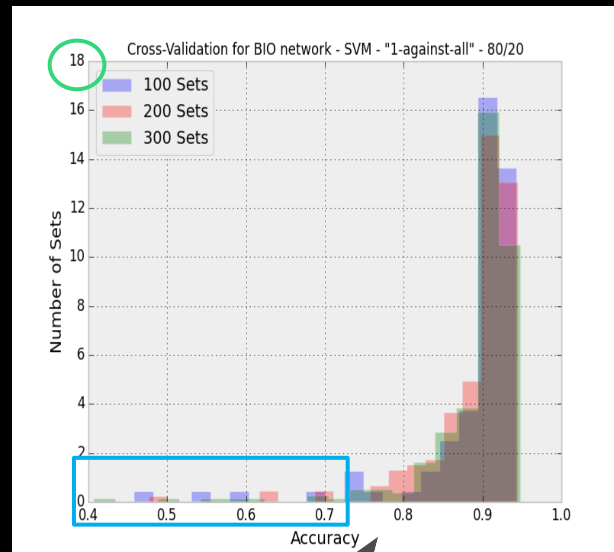
If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected!

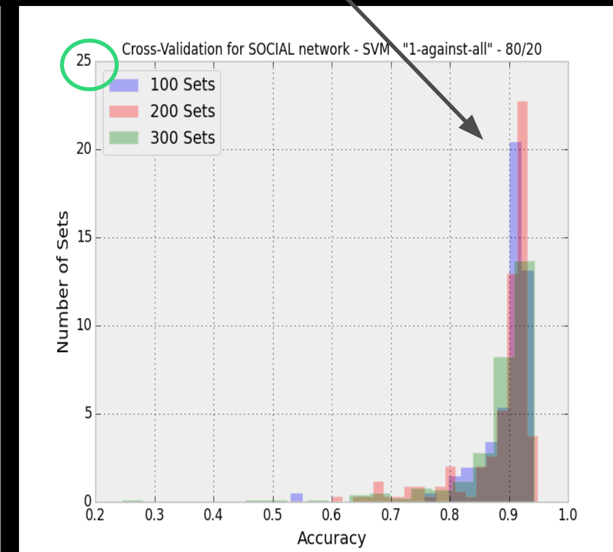


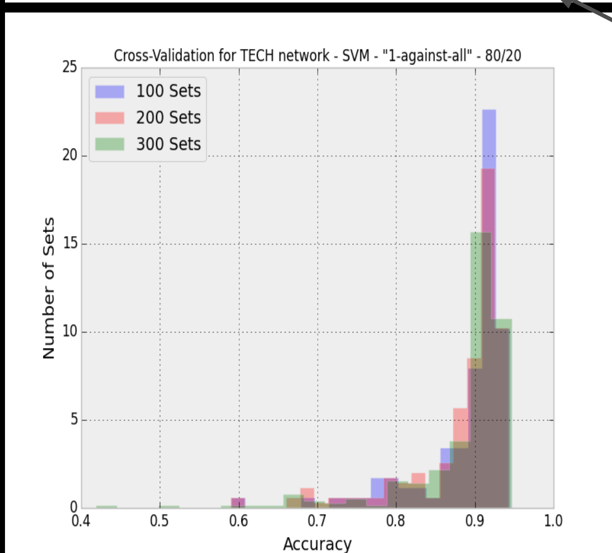
Cross Validation

social classifies better





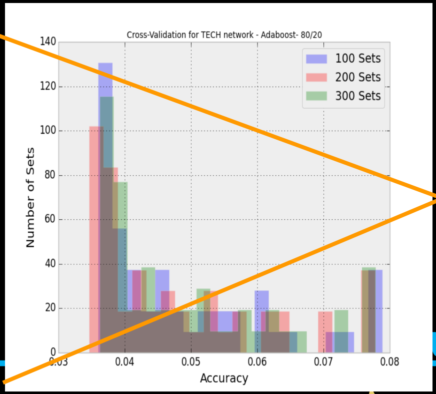


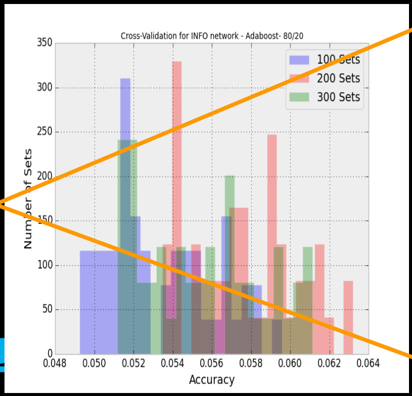


bio and tech classify worse

- **Smaller Samples Classify worse!**
- **We use to select good C (penalty parameter of the error term) and gamma (kernel coefficient.**

Outline

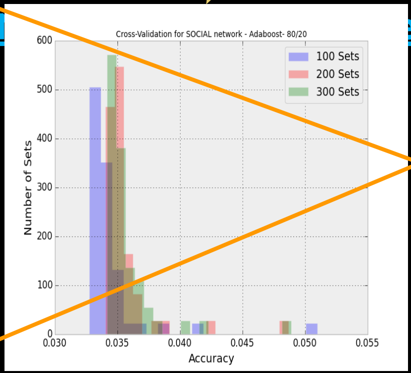


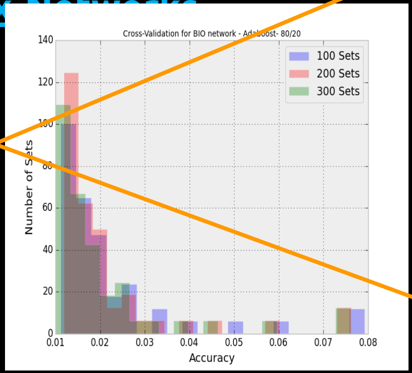


- **The classification seems to work :** u

- If you are interested: **astro.sunysb.edu/steinkirch/new/mloutputs.html**
  - More plots and results for:
    - Samplings
    - Cross-validation
    - Feature Selection: which of the 23 are really important?
    - Other supervisioned learning classifiers (**Adaboost**, LR, NB,...)
    - Some unsupervised learning

- You can try yourself with:
  - my results: **astro.sunysb.edu/steinkirch/new/mloutputs_sampled_tables/**
  - my code: **github.com/mariwahl/ML**





Thank you!