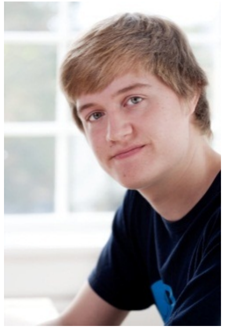
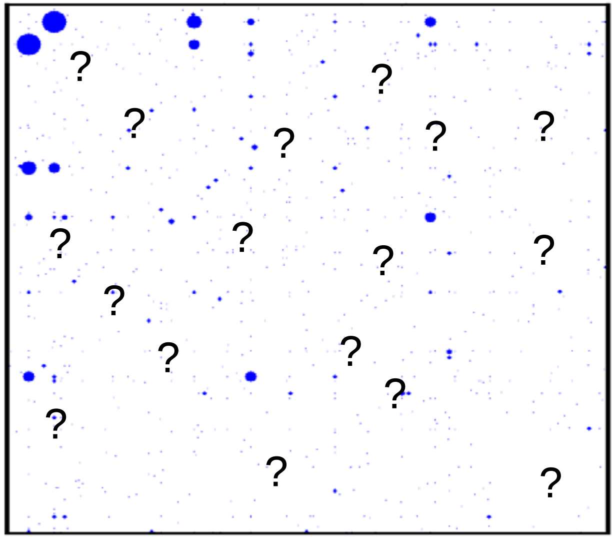
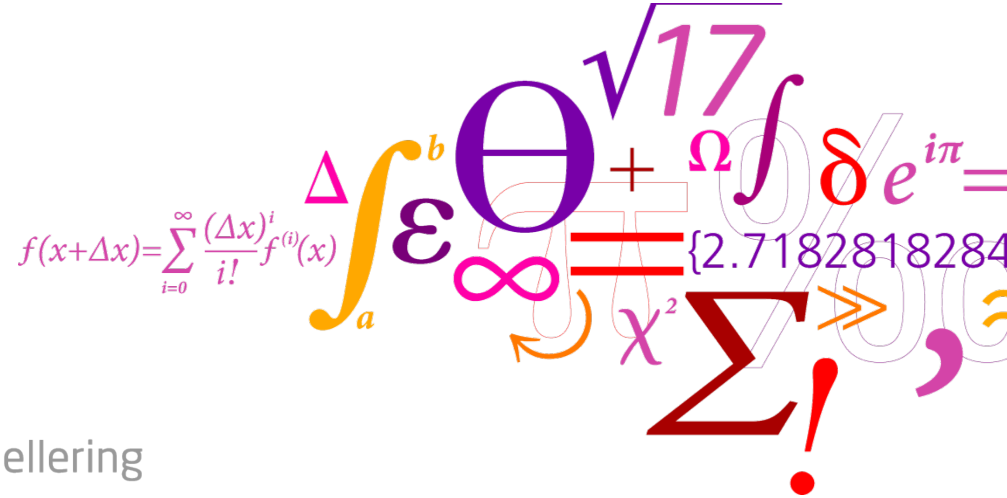
Link Prediction in Weighted Networks



David Kofoed Wind   Morten Mørup



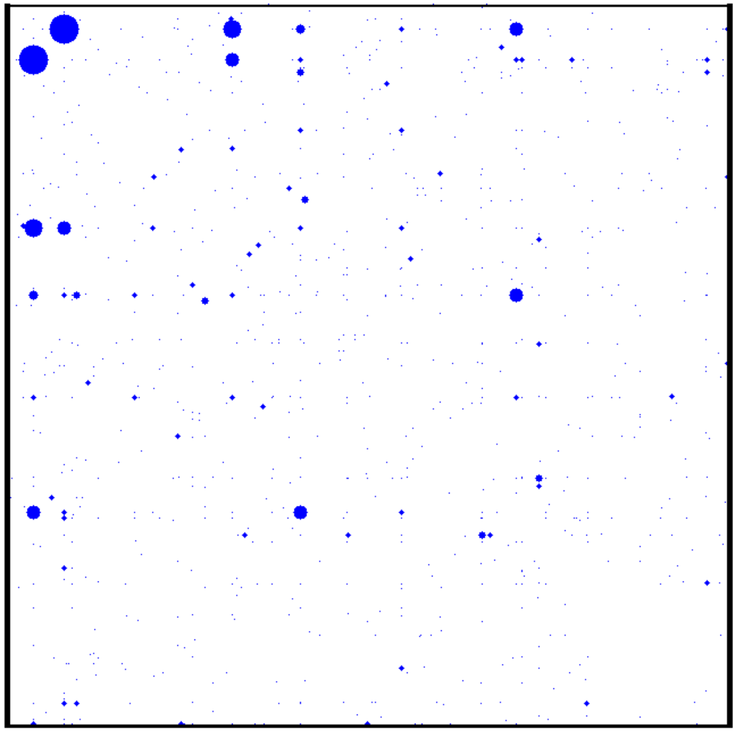


DTU Informatik

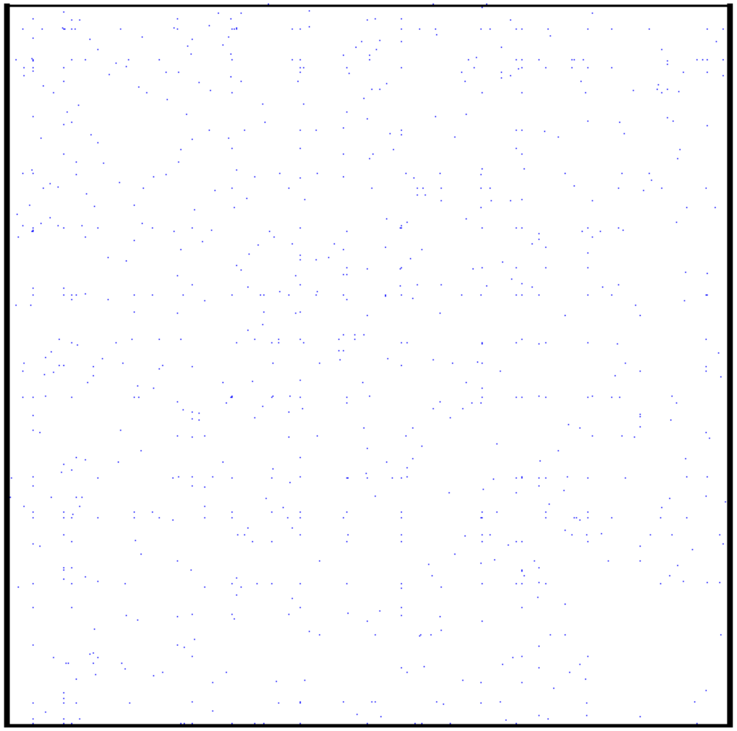
Institut for Informatik og Matematisk Modellering

Binary vs. Integer weighted networks

Integer Weighted Undirected Graph



Binary Undirected Graph
(Disregarding weight information)

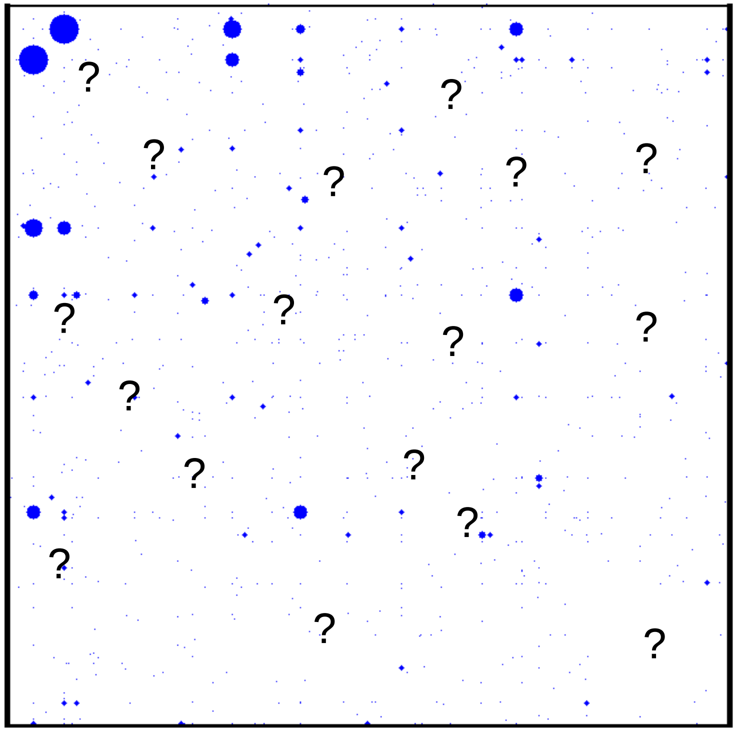


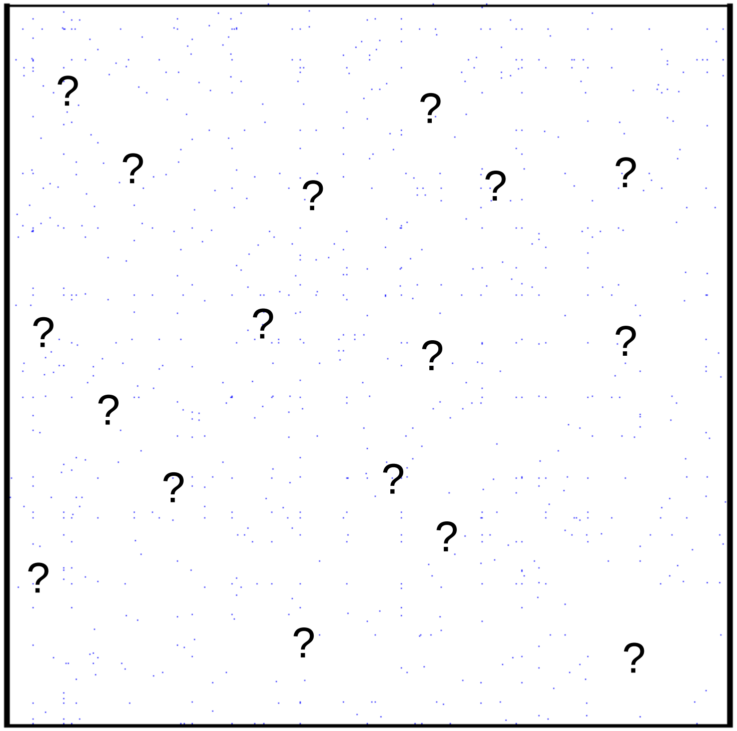
Does weight carry important information for the modeling of structure in networks?

The link prediction problem

1. Predict where a non-zero weight will become active by treating a set of links as zeros (see also Liben-Nowell and Kleinberg 2007, Clausset and Newman 2008).
   ->Realistic problem in networks in order to predict where links will emerge, straightforward to implement, admit comparison to non-statistical models that returns a link scoring function.
2. Predict links and non-links based on treating sets of links and non-links as missing in the model estimation (see for instance, Miller et al. 2010).
   ->Principled framework for predictive modeling, can be implemented using marginalization or imputation.

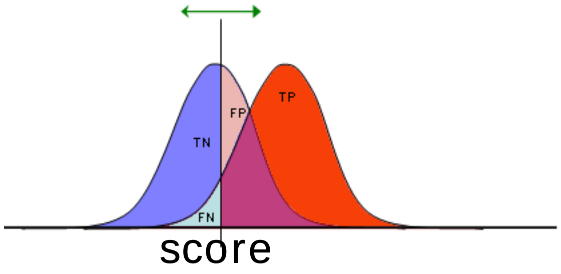
We will use the first approach (treating ?=0).



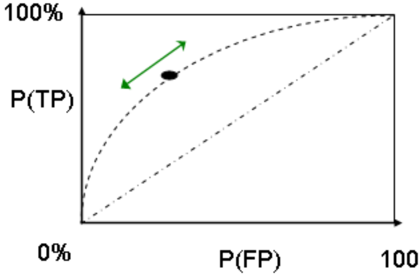


Predictive performance quantified by AUC

- Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) on hold out data

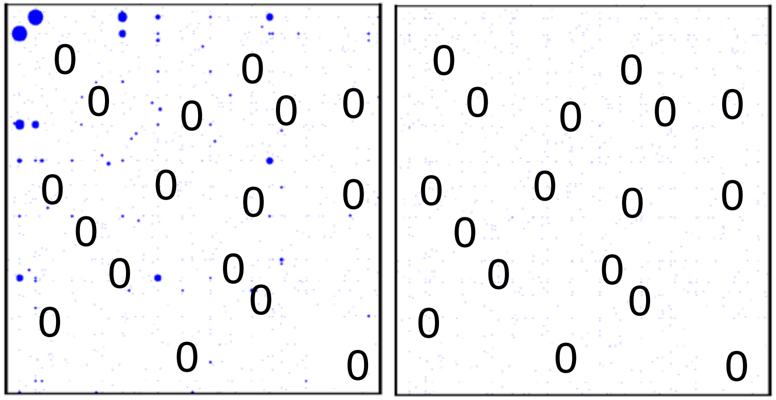


TP	FP
FN	TN
1	1



Benefit of the AUC is that it is not influenced by class-imbalance issues, i.e. in networks #links<<”#non-links

We will treat from 1 % to 99 % of the links as missing (i.e. unobserved) by setting them to zero and use AUC to quantify ability to account for links.

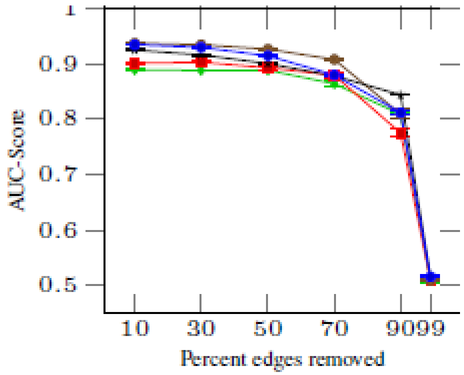


SBM: $\text{score}(i,j) = \eta_{z_i z_j}$

DCSBM: $\text{score}(i,j) = \theta_i \eta \theta_j$

NMF: $\text{score}(i,j) = \boldsymbol{w}_{i:} \boldsymbol{h}_j$

symNMF: $\text{score}(i,j) = \boldsymbol{w}_{i:} \boldsymbol{w}_j$



Models of graphs

- Stochastic Block Model
  (Holland et al. 1983, Snijders and Nowikcki 1997, Nowicki and Snijders 2001)

$$A_{ij} \sim \mathit{Bernoulli}(\eta_{z_i z_j}) = \eta_{z_i z_j}^{A_{ij}} (1-\eta_{z_i z_j})^{A^{ij}}$$

$$A_{ij} \sim \mathit{Poisson}(\eta_{z_i z_j}) = \frac{\eta_{z_i z_j}^{A_{ij}}}{A_{ij}!} \exp(-\eta_{z_i z_j})$$

- Degree Corrected Stochastic Block Model (Karreer and Newman 2011)

$$A_{ij} \sim \mathit{Poisson}(\theta_i \eta_{z_i z_j} \theta_j) = \frac{(\theta_i \eta \theta_j)_{z_i z_j}^{A_{ij}}}{A_{ij}!} \exp(-\theta_i \eta_{z_i z_j} \theta_j)$$

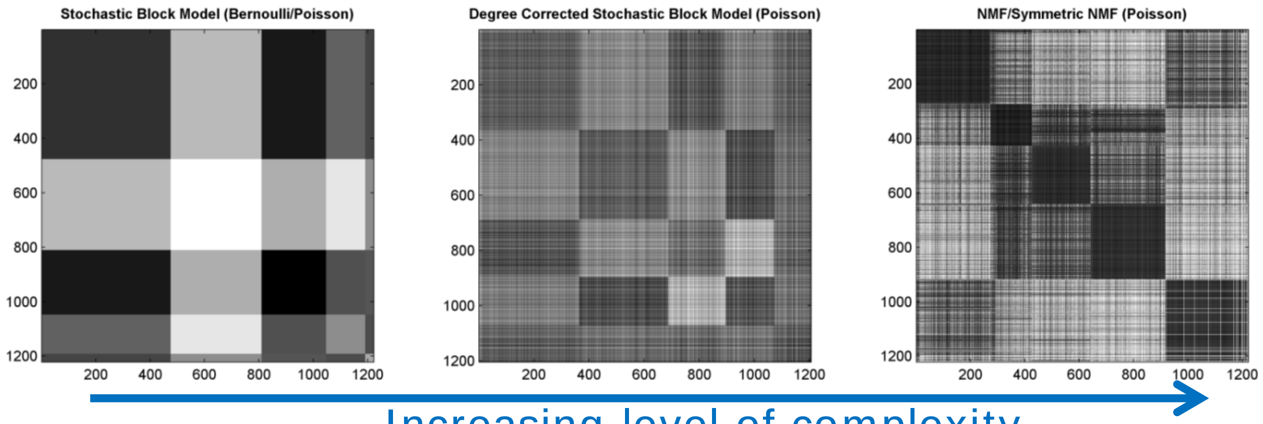
- Non-negative matrix factorization and symmetric NMF
  (Lee and Seung2001, Ding et al, 2005, Greeene et al. 2008 , Wang et al. 2010, Ball et al., 2011)

$$A_{ij} \sim \mathit{Poisson}(\boldsymbol{w}_{i:}\boldsymbol{h}_j) = \frac{(\boldsymbol{w}_{i:}\boldsymbol{h}_j)^{A_{ij}}}{A_{ij}!} \exp(-\boldsymbol{w}_{i:}\boldsymbol{h}_j)$$

$$A_{ij} \sim \mathit{Poisson}(\boldsymbol{w}_{i:}\boldsymbol{w}_j) = \frac{(\boldsymbol{w}_{i:}\boldsymbol{w}_j)^{A_{ij}}}{A_{ij}!} \exp(-\boldsymbol{w}_{i:}\boldsymbol{w}_j)$$

Inference by Stochastic search (See Karrer and Newman 2011)

Inference by Accelerated Multiplicative Updates (For details see paper)



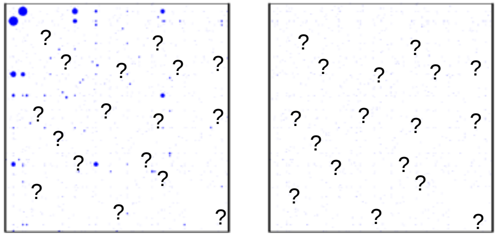
Increasing level of complexity

Research questions:

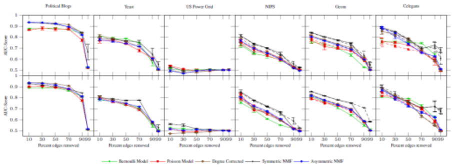
- Are binary graphs better modeled by Bernoulli than Poisson likelihood?

$$A_{ij} \sim \mathit{Bernoulli}(\eta_{z_iz_j}) = \eta_{z_iz_j}^{A_{ij}}(1-\eta_{z_iz_j})^{A^{ij}} \quad \text{vs.} \quad A_{ij} \sim \mathit{Poisson}(\eta_{z_iz_j}) = \frac{\eta_{z_iz_j}^{A_{ij}}}{A_{ij}!}\exp(-\eta_{z_iz_j})$$

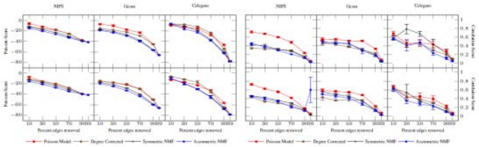
- Are edge-weights useful for the prediction of links?



- Are detection thresholds influenced by model complexity?



- Are predictions of edge-weight influenced by model complexity?



All questions are adressed by link prediction following the framework of (Liben-Nowell and Kleinberg 2007, Clausset and Newman 2008).

Networks considered

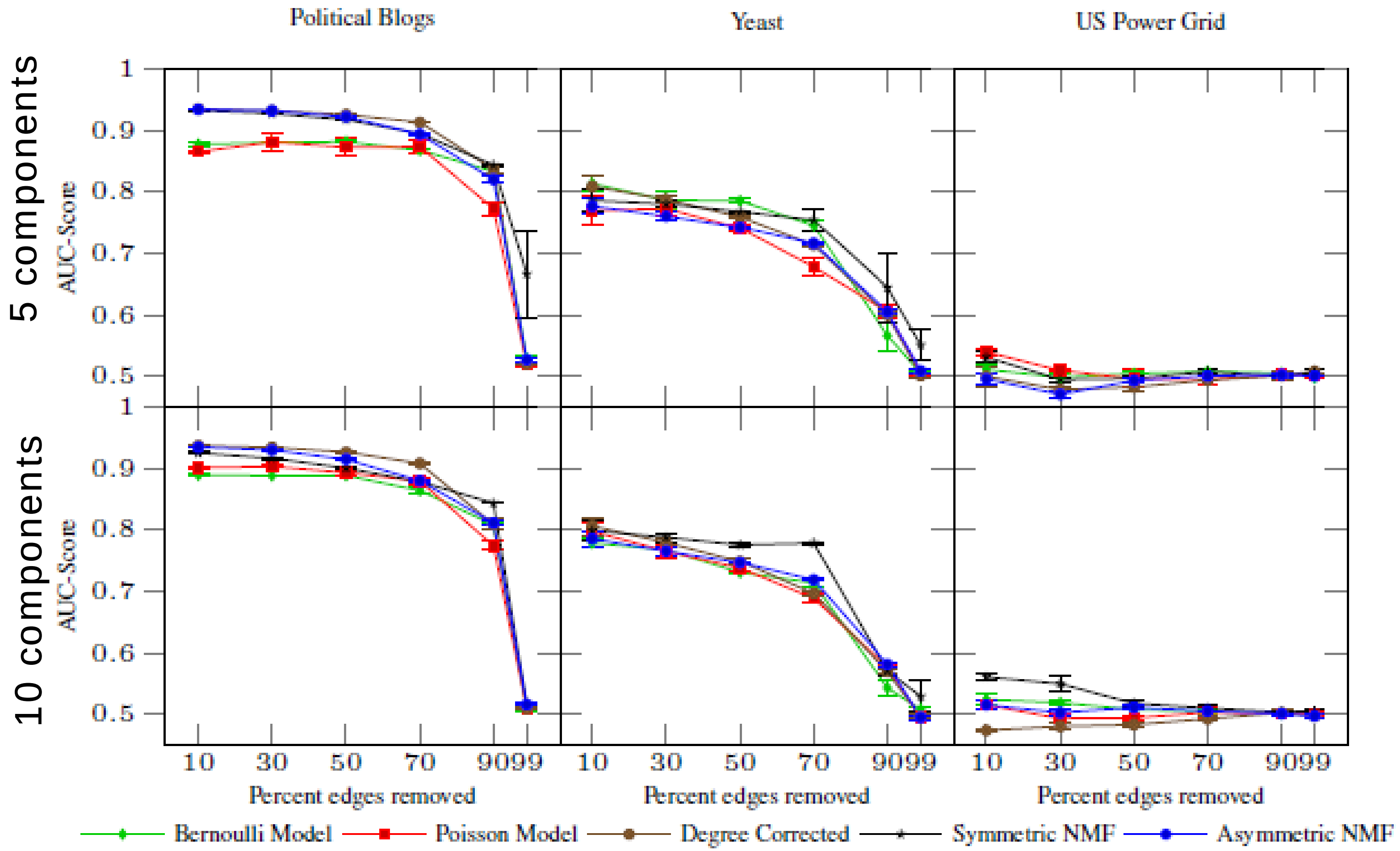
Binary

- **Political Blogs:** The Political Blogs-network is assembled by (Adamic and Glance, 2005). The nodes are political blogs about US politics, and the edges are links between them, all captured on a single day in 2005. We only consider the largest connected component containing 1222 vertices, and 16718 edges containing no weight information.
- **Yeast:** The Yeast network is assembled by (Sun et. al., 2003). The vertices represent proteins in budding yeast, and the edges interactions between them. The network has 2361 vertices, and 6647 edges, and the edges contains no weight information.
- **US Power Grid:** The US Power Grid network (Tsaparas, 2006), models the structure of the power grid in Western US. The network has 4941 vertices, 6595 edges and contains no weight information.

Integer Weighted

- **NIPS:** The NIPS-network is a co-authorship dataset assembled by (Globerson et. al., 2005), containing information on publications of the NIPS-conferences from 1988 to 2003. Vertices corresponds to the authors, and edges to joint publications. The weight of an edge is the number of joint publications by the authors. The network has 2865 vertices and 4734 edges.
- **Geom:** The Geom network (Batagelj, 2006) is a collaboration network from computational geometry. Vertices corresponds to the authors, and edges to joint publications. The weight of an edge is the number of joint publications by the authors. The network contains 7343 vertices and 11898 edges.
- **Metabolic pathways:** The metabolic pathways network is an integer weighted network of metabolic pathways in 43 organisms (Jeong, et al. 2000), where vertices are the substrates, and edges are the metabolic reactions. The weight of an edge, corresponds to multiple metabolic reactions between a pair of substrates. The network has 453 vertices, and 2026 edges.

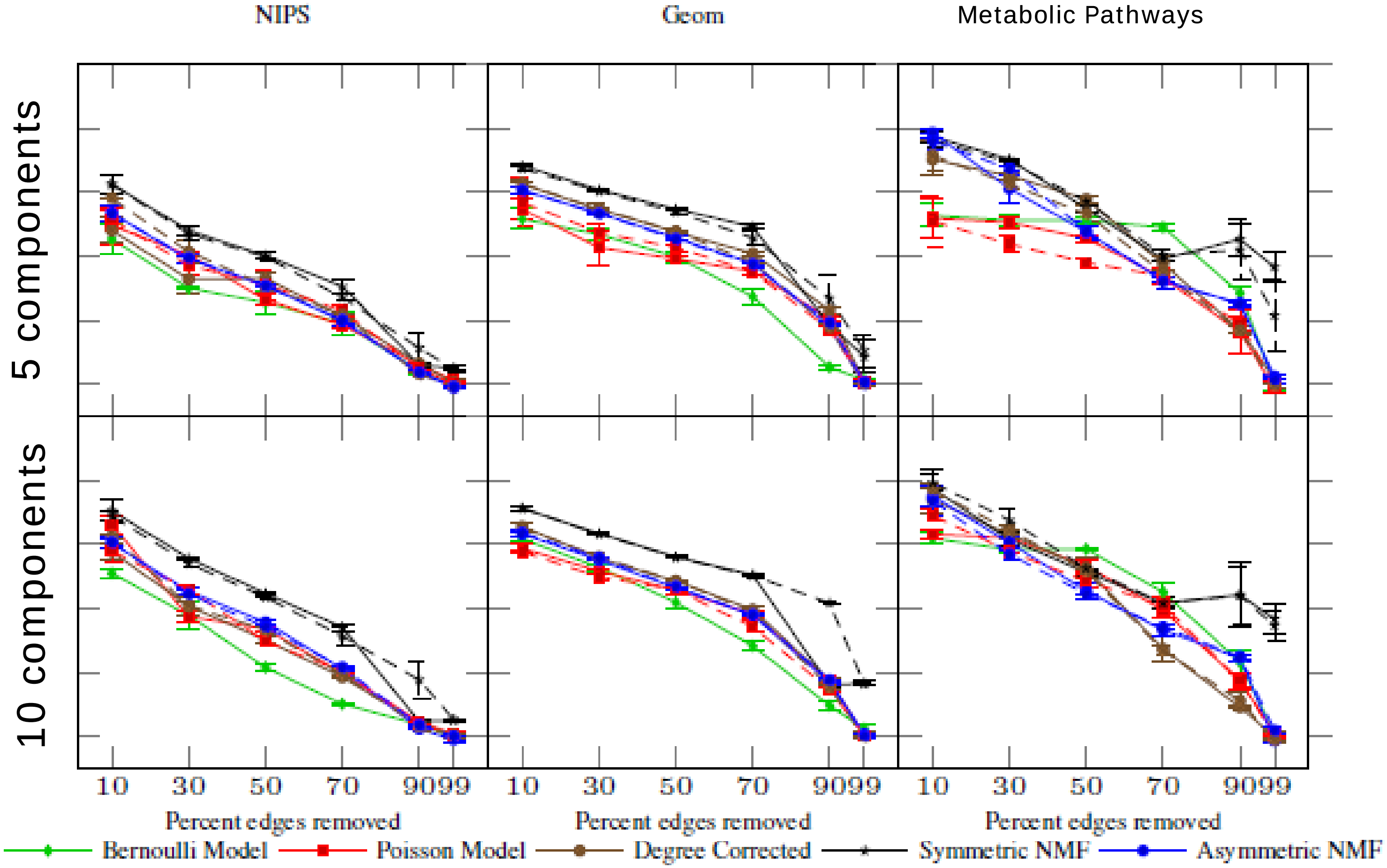
Are binary graphs better modeled by Bernoulli than Poisson likelihood?



No it doesn’t seem to be the case. This is also expected as the Bernoulli distribution can be approximated by Poisson distribution (se paper for details).

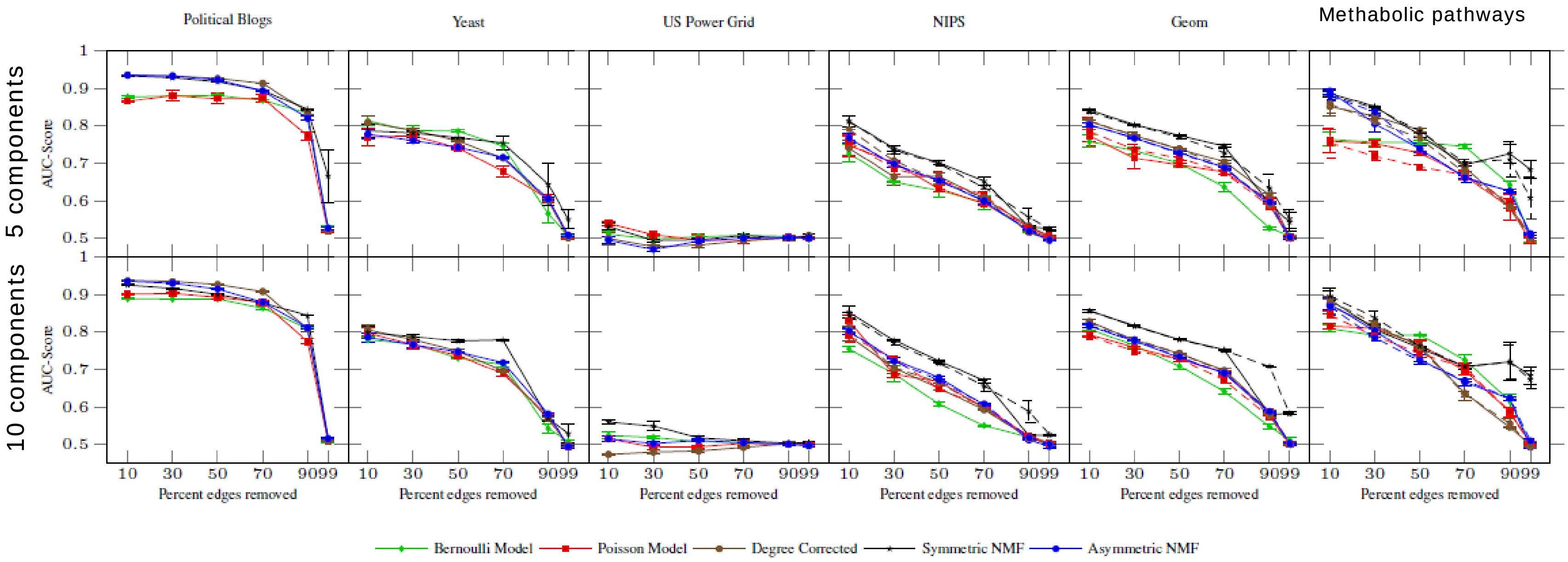
Are edge-weights useful for the prediction of links?

Solid: Integer Weighted graph
Dashed: Corresponding Binary graph



No, in general it does not appear to be the case that weight information improve on the prediction of links.

Are detection thresholds influenced by model complexity?



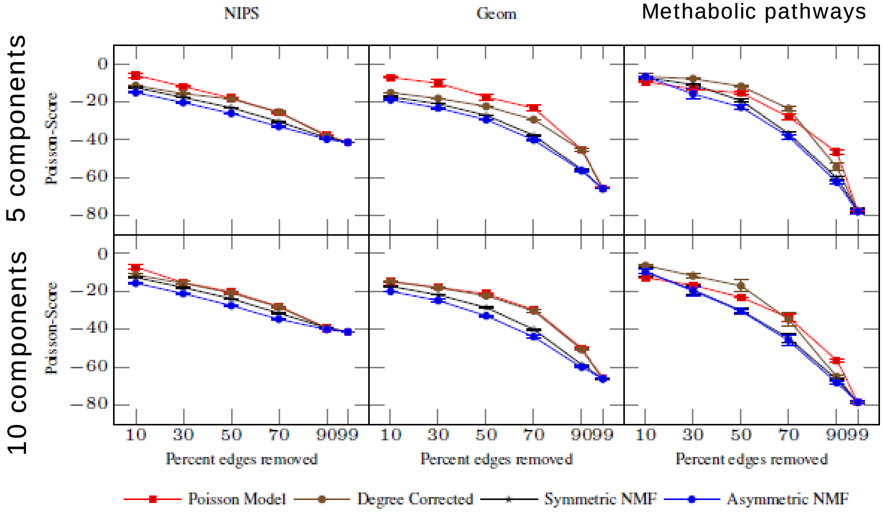
No, It appears that complex models in general outperform simpler models regardless of the fraction of links observed in the networks.

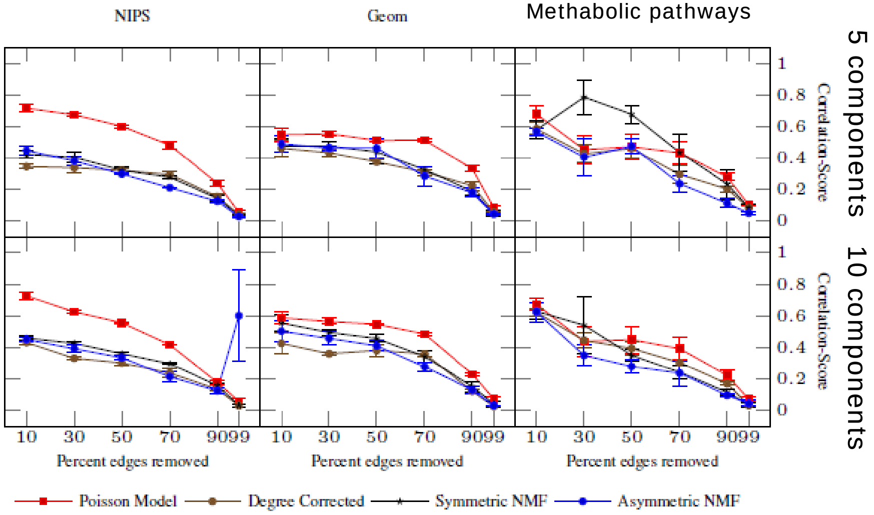
Are predictions of edge-weight influenced by model complexity?

Predicting weight values by a Poisson scoring function as well as by correlation.

$$S=\log\prod_{(ij)\in\mathcal{M}}\frac{r_{ij}^{A_{ij}}}{A_{ij}!}e^{-r_{ij}}=\sum_{(ij)\in\mathcal{M}}A_{ij}\log(r_{ij})-\Gamma(A_{ij}+1)-r_{ij}$$

$$S=\frac{\sum_{(i,j)\in\mathcal{M}}(A_{ij}-\bar{A})(r_{ij}-\bar{r})}{\sqrt{\left(\sum_{i,j}(A_{ij}-\bar{A})^2\right)\left(\sum_{i,j}(r_{ij}-\bar{r})^2\right)}}$$

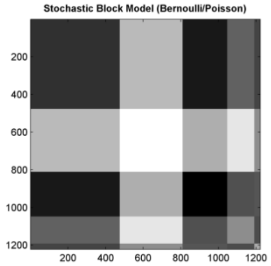


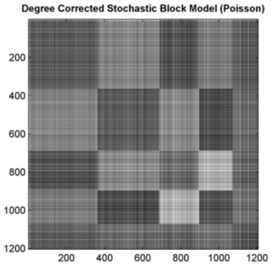


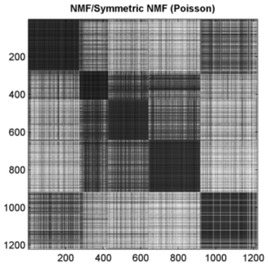
Yes, it appears that complex models in general overfits to the links treated as zeros, thus complex models are less able to recover edge weight compared to simpler less flexible models.

Discussion

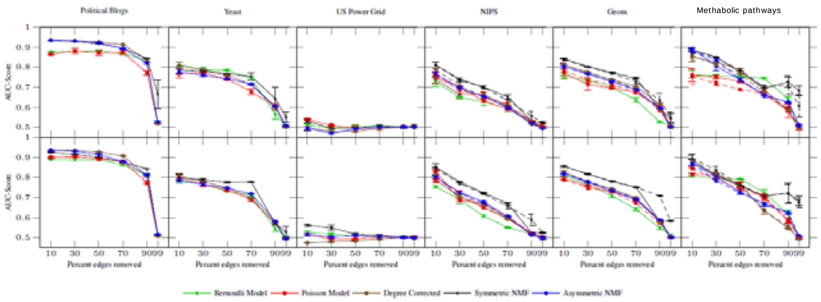
- **We considered four models of increasing complexity**
  Stochastic Block Model
  Degree Stochastic Block Model
  Non-negative Matrix Factorization (NMF)
  Symmetric NMF

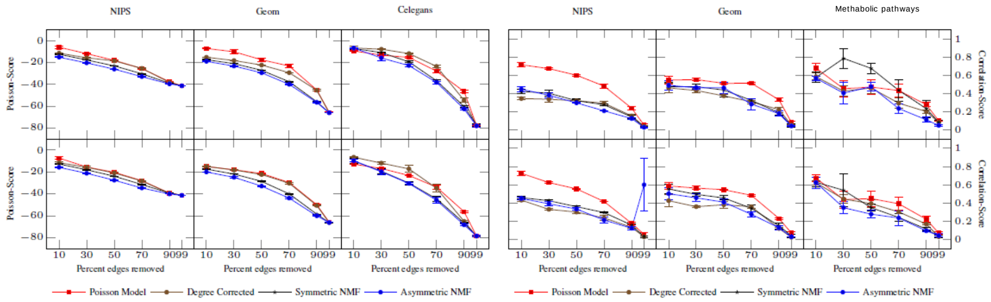






- **We applied link prediction treating links as zeros.**
- **We considered the following four questions analyzing six real networks**
  Are binary graphs better modeled by Bernoulli than Poisson likelihood? Not really
  Are edge-weights useful for the prediction of links? Not really
  Are detection thresholds influenced by model complexity? Not in our experiments, complex models outperform simpler models.
  Are predictions of edge-weight influenced by model complexity? Yes it appears so.


Political Blogs
Yeast
US Power Grid
NIPS
Geom
Methabolic pathways
AUC-Score
Percent edges removed
Bernoulli Model
Poisson Model
Degree Corrected
Symmetric NMF
Asymmetric NMF


NIPS
Geom
Celegans
NIPS
Geom
Methabolic pathways
Poisson-Score
Correlation-Score
Percent edges removed
Poisson Model
Degree Corrected
Symmetric NMF
Asymmetric NMF