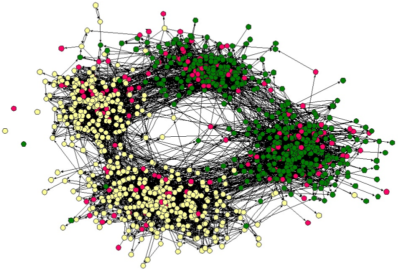
Hypothesis- vs. Data-Driven Research

Jörg Reichardt
-
reichardt@physik.uni-wuerzburg.de

Institute for Theoretical Physics, University of Würzburg, Germany
joint work with
Michele Leone, ISI Torino, Italy

Zürich, August 20, 2008

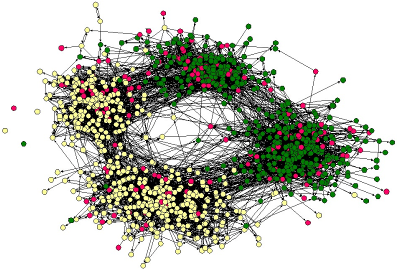
Hypothesis Driven



- Needs hypothesis
- Needs appropriate data (interactions+properties)
- Statistical Significance: p-value
- Small effects seen in lots of data

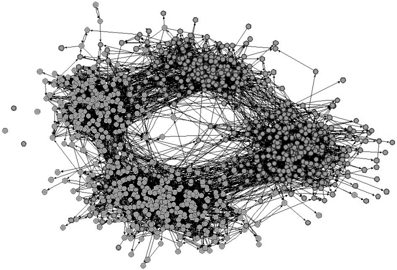
(picture by Mark Newman)

Hypothesis Driven



- Needs hypothesis
- Needs appropriate data (interactions+properties)
- Statistical Significance: p-value
- Small effects seen in lots of data

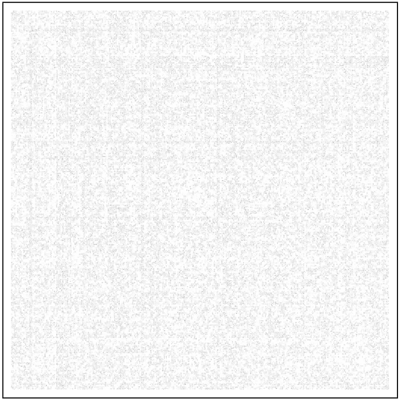
Data Driven



- No Hypothesis needed
- No full data needed (only interactions)
- Post-hoc explanation
- Statistical Significance?!
- Effect size?!

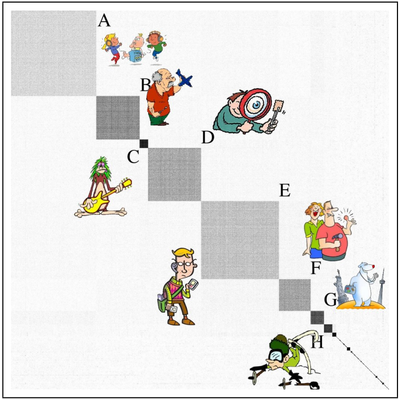
(picture by Mark Newman)

Market Research as an Example

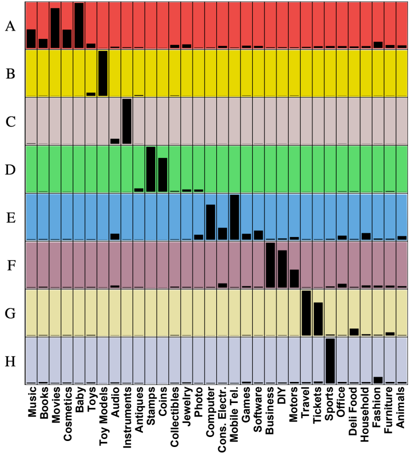


- $N = 892,641$ eBay users
- $M = 7,4$ Mio links (pairwise competitions for single articles)
- Infer possible hidden classes of agents (interest groups)
- Reorder rows and columns according to classes

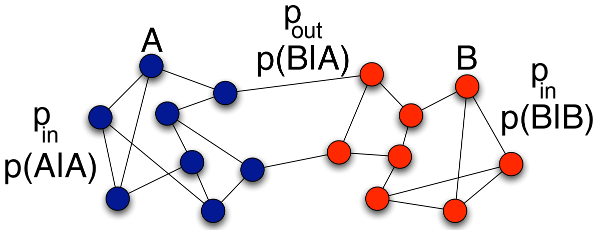
Interpretation of Bidder Groups



Risk ratio of bidding in category

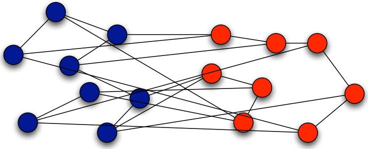


A well defined Problem: Planted Partitions

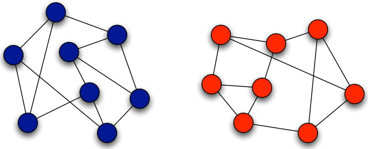


- Ensemble of (infinitely) large Network with given $p(k)$ and $\sum_k^{\infty} kp(k) = \langle k \rangle$ **finite**
- Nodes carry hidden cluster index $s_i \in \{1, 2\}$ (type A,B).
- Wiring is random except for within/between group wiring
- One parameter: a fraction of p_{in} links lies within clusters, the rest between clusters (equal sized for simplicity).
- **Can we infer the colors given links, sizes and number of clusters, only?**

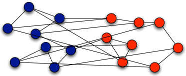
Impossible-to-Trivial-Transition



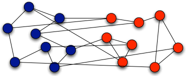
impossible for $p_{in} = 0.5$



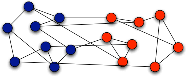
trivial for $p_{in} = 1.0$



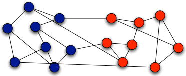
$p_{in} = 0.58$



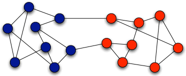
$p_{in} = 0.66$



$p_{in} = 0.75$

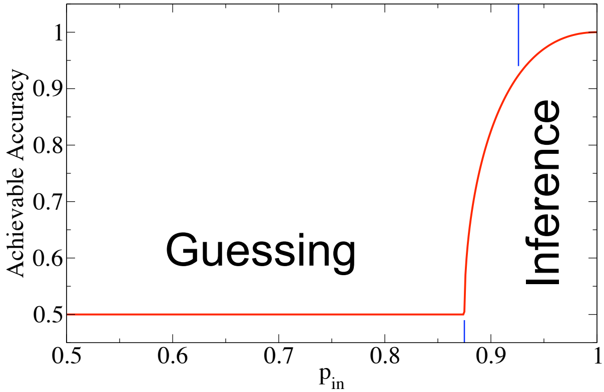


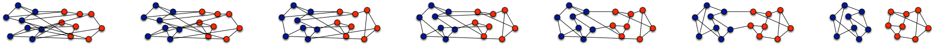
$p_{in} = 0.83$



$p_{in} = 0.92$

A Worst Case Scenario: 3 Links per Node





0.5 0.58 0.66 0.75 0.83 0.92 1.0

Why this transition?

- Given only the network A_{ij}, size and number of clusters
- Only sensible approach: Look for maximally separated clusters!
- Find a minimum cut, i.e. find the ground state (global minimum) of:

$$\text{Cutsize } E = \sum_{i<j} A_{ij}(1 - \delta(\sigma_i, \sigma_j))$$

under constraint $\frac{1}{N} \sum_i \delta(\sigma_i, r) = 1/2$ for all $r \in \{1, 2\}$

- Effectively: among all $N!/(N/2)!/(N/2)!$ partitions into two equal sized clusters ("configurations"), find the one with minimum number of edges between clusters (Bayes MAP optimal)
- Note: Cutsize of planted cluster structure: $E^p = N\frac{\langle k \rangle}{2}(1 - p_{in})$

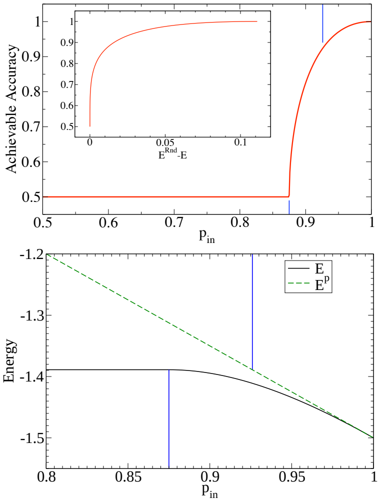
Algorithm Independent Results

- **Problem:** Designed configuration is a guaranteed local minimum of the cutsize only (!) for $p_{in} = 1$.
- **Study the overlap of the expected configuration which minimizes E with planted clusters as function of p_{in}.**
- Makes analysis independent of inference algorithm used and results universal.
- Statistical Physics allows to calculate $p(\sigma_i|s_i)$ as a function of p_{in}
- Find the expected accuracy of recovering the hidden variables via

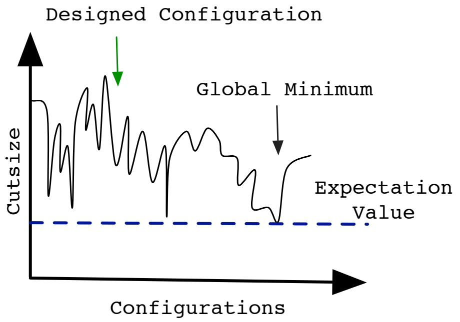
$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\delta(\sigma_i, s_i) = \sum_{s} p(\sigma = s|s)$$

where the σ_i minimize the cutsize E and s_i are the hidden variables.

Influence of Graph Topology on Min-Cut Partition

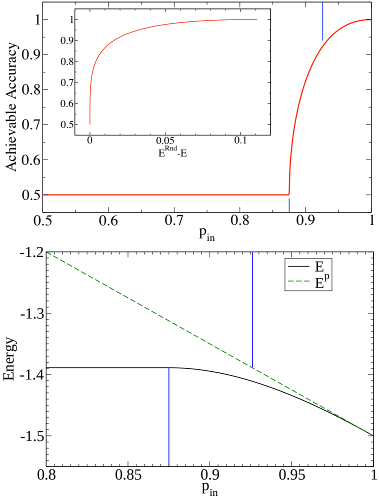


J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)

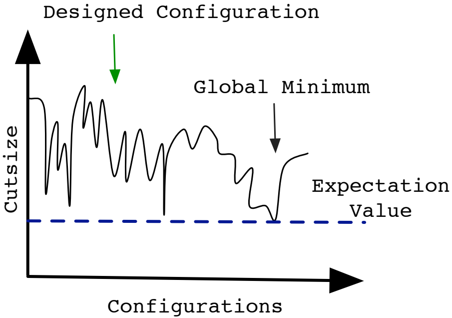


- Can find small cutsizes even in random networks
- Alternative minima compete with designed minima

Influence of Graph Topology on Min-Cut Partition

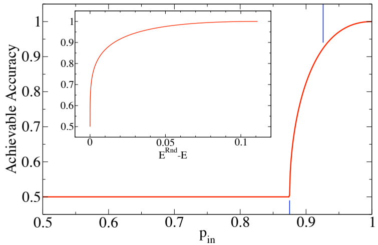


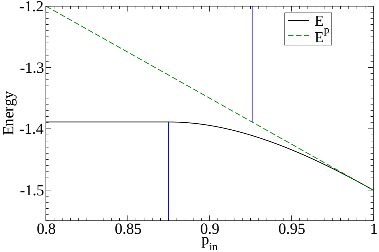
J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)



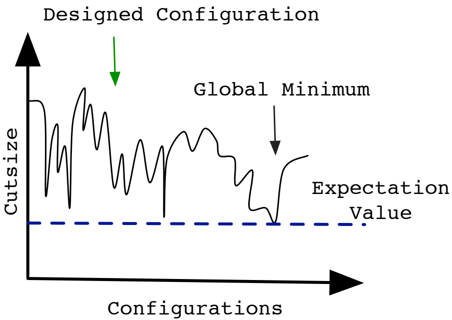
- Can find small cutsizes even in random networks
- Alternative minima compete with designed minima

Influence of Graph Topology on Min-Cut Partition



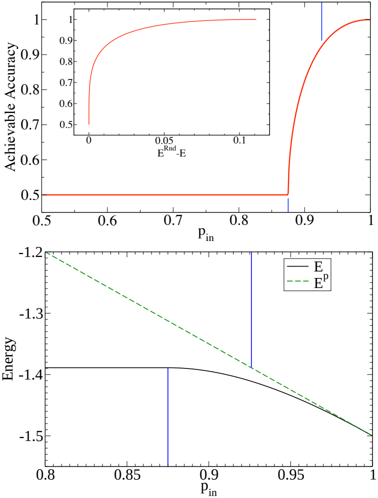


J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)

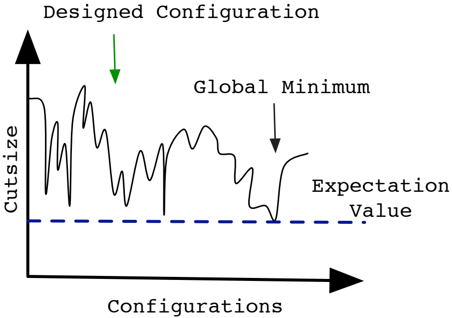


- Can find small cutsizes even in random networks
- Alternative minima compete with designed minima

Influence of Graph Topology on Min-Cut Partition

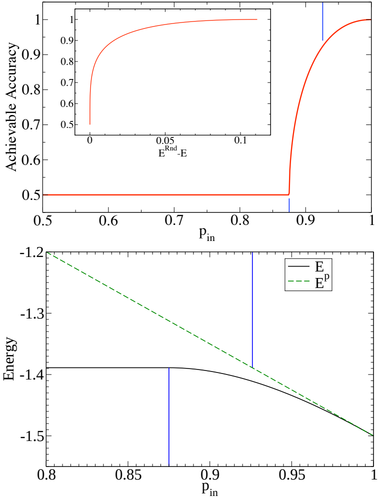


J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)

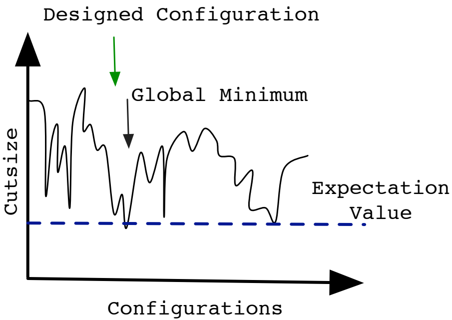


- Can find small cutsizes even in random networks
- Alternative minima compete with designed minima

Influence of Graph Topology on Min-Cut Partition

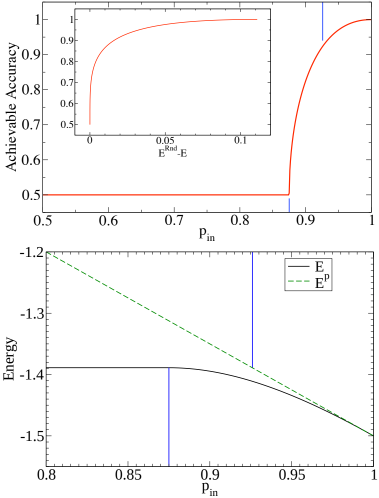


J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)

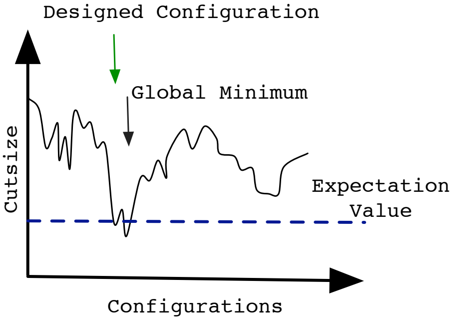


- At $p_{in} \geq p_{in}^c$ we find a configuration that has a lower energy than expected in a random network.
- This global minimum moves closer to the designed configuration

Influence of Graph Topology on Min-Cut Partition

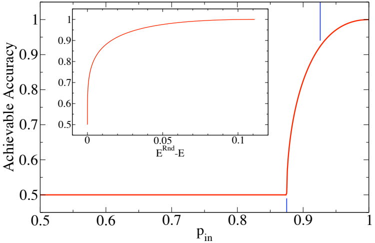


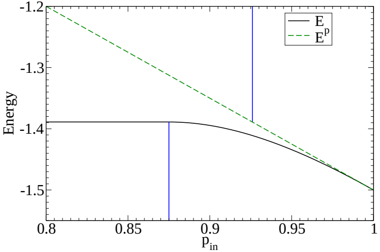
J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)



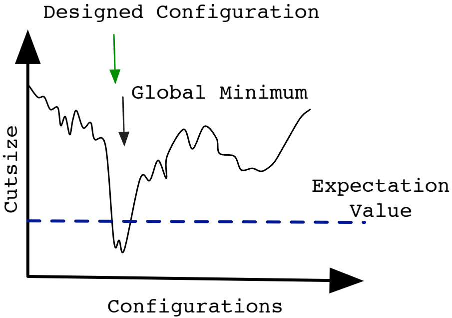
- At $p_{in} = 2E^{Rnd}/\langle k \rangle$ the designed minimum is lower than the expectation value in a random network
- Less local minima

Influence of Graph Topology on Min-Cut Partition



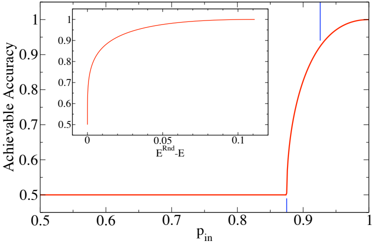


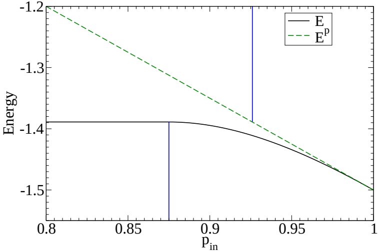
J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)



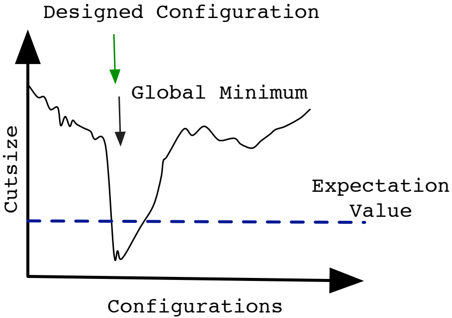
- Global minimum approaches designed configuration with increasing p_{in}
- Less local minima, landscape smoothes

Influence of Graph Topology on Min-Cut Partition



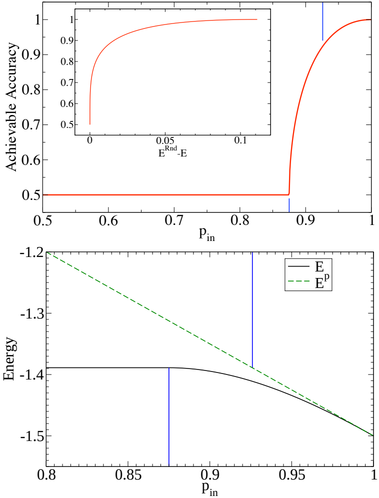


J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)

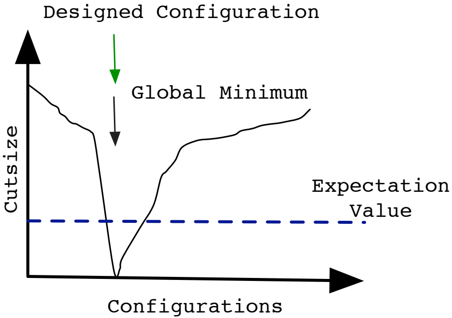


- Global minimum approaches designed configuration with increasing p_{in}
- Less local minima, landscape smoothes

Influence of Graph Topology on Min-Cut Partition

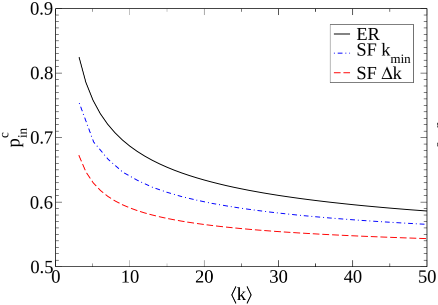


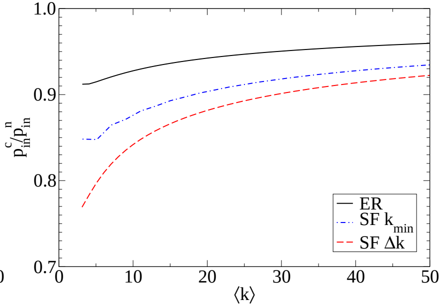
J.R.,M. Leone, Phys. Rev. Lett, **101**, 078701 (2008)



- At $p_{in} = 1$ designed minimum and global minimum coincide
- Only one minimum left

How does p_{in}^c depend on Degree Distribution?

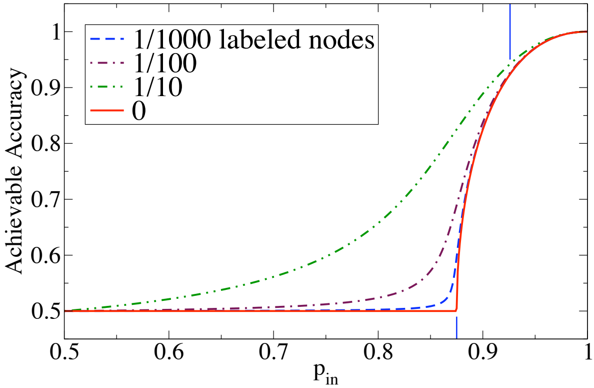




- ER: Poissonian, SF k_{min}: $p(k) \propto k^{-3}$ for $k \geq k_{min}$,
  SF Δk: $p(k) \propto (k + \Delta k)^{-3}$
- Naïve guess for critical p_{in} would be $p_{in}^n = 2E^{Rnd}/\langle k \rangle$ and is too conservative.
- Recognizable structure starts at "weaker" cluster structures.

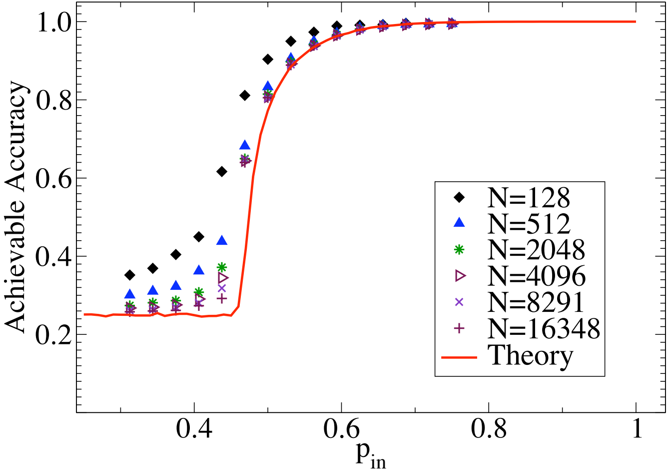
Inclusion of Prior Knowledge

Again, only 3 links per node, finite fraction of hidden labels known:



- Partially labeld data may increase accuracy dramatically
- Especially around the transition point.

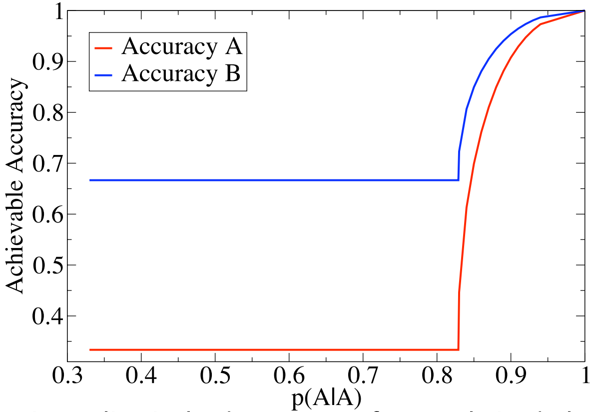
Finite Size Effects



4 equal sized groups, Poissonian $p(k)$ with $\langle k \rangle = 16$

Unequal Cluster Sizes

Bethe lattice with 3 links per node, 2/3 type A, 1/3 type B



- Behavior is qualitatively the same as for equal sized clusters
- Transition point changes slightly (p_{in}^{c} moves left)

Conclusion

- Sharp transition from impossible to easy cluster detection

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!
- "Dimensionality and size of data set are not independent".

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!
- "Dimensionality and size of data set are not independent".
- There may exist structure that is principally undetectable by unsupervised methods even in infinitely large networks.

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!
- "Dimensionality and size of data set are not independent".
- There may exist structure that is principally undetectable by unsupervised methods even in infinitely large networks.
- Spurious solutions in large "hypothesis space" obscure true structure.

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!
- "Dimensionality and size of data set are not independent".
- There may exist structure that is principally undetectable by unsupervised methods even in infinitely large networks.
- Spurious solutions in large "hypothesis space" obscure true structure.
- Inclusion of prior knowledge (labeled nodes) may help somewhat.

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
  - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
  - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!
- "Dimensionality and size of data set are not independent".
- There may exist structure that is principally undetectable by unsupervised methods even in infinitely large networks.
- Spurious solutions in large "hypothesis space" obscure true structure.
- Inclusion of prior knowledge (labeled nodes) may help somewhat.
- Analytical formulae for transition point and achievable accuracy.

Conclusion

- Sharp transition from impossible to easy cluster detection
- Similar transitions for multivariate data:
    - Given $N = \alpha D$ data points in a space of dimension D, can we infer clusters (Gaussian Mixtures, etc)?
    - Answer: Yes we can, if only $\alpha > \alpha_c$! (Given enough data, we can learn any distribution)
- This is wrong for sparse graphs (those with finite connectivity)!
- "Dimensionality and size of data set are not independent".
- There may exist structure that is principally undetectable by unsupervised methods even in infinitely large networks.
- Spurious solutions in large "hypothesis space" obscure true structure.
- Inclusion of prior knowledge (labeled nodes) may help somewhat.
- Analytical formulae for transition point and achievable accuracy.

Data driven research will (only) tell you about (all) strong effects!
Small effects are visible only to hypothesis driven research!

References

- J.R. and S. Bornholdt: Clustering of sparse data via network communities - a prototype study of a large online market J. Stat. Mech. (2007) P06016
- J.R. and S. Bornholdt: Partitioning and modularity of graphs with arbitrary degree distribution, Phys. Rev. E **76**, 015102(R) (2007)
- J.R. and M. Leone: (Un)detectable cluster structure in sparse networks, Phys. Rev. Lett. **101**, 078701 (2008)
- J.R. and D.R. White: Role Models for Complex Networks, Eur. Phys. J. B **60**, 217-224 (2007)

Solution via Cavity-Equations

$$
\begin{aligned}
P(\mathbf{h}|s) &= \sum_{k=0}^{\infty} p(k) \int \prod_{i=1}^{k} \left(d^q \mathbf{u}_i Q_{in}(\mathbf{u}_i|s)\right) \delta \left(\mathbf{h} - \sum_{i=1}^{k} \mathbf{u}_i \right) \\
Q(\mathbf{u}|s) &= \sum_{d=0}^{\infty} q(d) \int \prod_{i=1}^{d} \left(d^q \mathbf{u}_i Q_{in}(\mathbf{u}_i|s)\right) \delta \left(\mathbf{u} - \hat{\mathbf{u}} \left(\sum_{i=1}^{d} \mathbf{u}_i \right) \right)
\end{aligned}
$$

$$
Q_{in}(\mathbf{u}|s) = p_{in} Q(\mathbf{u}|s) + \sum_{r \neq s}^{q} \frac{1 - p_{in}}{q - 1} Q(\mathbf{u}|r).
$$

$\underline{Q(\mathbf{u}|s) = \eta_{cw}, \text{ where } c = u^s \text{ and } w = ||\mathbf{u}||^2 - c}$

Symmetry considerations enforce equi-partition and reduce the number of independent parameters from $q(2^q - 1)$ to only $2q - 1$!

Iterated Solution of Cavity-Equations for 2 Clusters

$$
\begin{aligned}
\eta_{11} &= \sum_{n0=0}^{\infty} \sum_{n=0}^{\infty} q(n_0+2n) \frac{(n_0+2n)!}{n_0!n!n!} \left(\eta_{10}^{in}\right)^n \left(\eta_{01}^{in}\right)^n \eta_{11}^{n_0} \\
\eta_{10} &= \sum_{n0=0}^{\infty} \sum_{n_1>n_2}^{\infty} q(n_0+n_1+n_2) \frac{(n_0+n_1+n_2)!}{n_0!n_1!n_2!} \left(\eta_{10}^{in}\right)^{n_1} \left(\eta_{01}^{in}\right)^{n_2} \eta_{11}^{n_0} \\
\eta_{01} &= 1-\eta_{11}-\eta_{10}
\end{aligned}
$$