NETWORK HYPOTHESIS TESTING USING MIXED KRONECKER PRODUCT GRAPH MODELS

Sebastian Moreno, Jennifer Neville
Computer Science Department
Purdue University

*Presented by Joel Pfeiffer*
Introduction

• Most research on networks has focused on analyzing the characteristics of a single large network drawn from a specific domain.

• Few statistical methods are available to determine if similarities observed across networks are expected, or if observed differences are significant.
Introduction

- Most research on networks has focused on analyzing the characteristics of a single large network drawn from a specific domain.

We propose a new model-based, across-network hypothesis testing framework to determine if two networks are sampled from the same underlying graph distribution.
General hypothesis testing

- One sample hypothesis tests have the same basic format:

\[ Y \quad X \sim N(?,?) \]
One sample hypothesis tests have the same basic format:
- Formulate null and alternative hypotheses ($H_0$ and $H_1$).

\[ H_0: Y \sim N(\bar{X}, s_X) \quad H_1: Y \not\sim N(\bar{X}, s_X) \]
One sample hypothesis tests have the same basic format:

- Formulate null and alternative hypotheses ($H_0$ and $H_1$).
- Identify statistic $T$ and determine the distribution $P(\cdot)$ under $H_0$. 

$$H_0: Y \sim N\left(\overline{X}, s_X \right) \quad H_1: Y \not\sim N\left(\overline{X}, s_X \right)$$

$$T = \overline{X}$$

$$Y \
\sim N\left(?, ? \right) \quad X \sim N\left(\overline{X}, s_X \right)$$
One sample hypothesis tests have the same basic format:

- Formulate null and alternative hypotheses (H_0 and H_1).
- Identify statistic T and determine the distribution P(\cdot) under H_0.
- Compute P(t_1 \leq T \leq t_2 | H_0) = 1 - \alpha based on H_0.
General hypothesis testing

• One sample hypothesis tests have the same basic format:
  – Formulate null and alternative hypotheses ($H_0$ and $H_1$).
  – Identify statistic $T$ and determine the distribution $P(\cdot)$ under $H_0$.
  – Compute $P(t_1 \leq T \leq t_2 | H_0) = 1 - \alpha$ based on $H_0$.
  – Calculate $t_{obs}$ the value of the statistic from the observed data.
General hypothesis testing

- One sample hypothesis tests have the same basic format:
  - Formulate null and alternative hypotheses (H₀ and H₁).
  - Identify statistic T and determine the distribution P(·) under H₀.
  - Compute P(t₁ ≤ T ≤ t₂|H₀)=1-α based on H₀.
  - Calculate t_{obs} the value of the statistic from the observed data.
  - If t_{obs}<t₁, or t_{obs}>t₂ the null hypothesis H₀ is rejected.
For network hypothesis testing, the underlying distribution of graphs $P(\cdot)$ is generally unknown.
Network hypothesis testing

• We propose to learn a model of the underlying distribution from an observed graph.
• Then we use the learned distribution to determine if a new network is likely to have been drawn from the same distribution.
Issues for network hypothesis testing

- Current statistical methods have some difficulties modeling the underlying graph distribution $P(\cdot)$
  - Errors in estimating the underlying probability distribution.

Incorrect estimate of variance (KPGM, CL, ERGM)

Biased estimate of mean
Issues for network hypothesis testing

- Current statistical methods have some difficulties modeling the underlying graph distribution \( P(\cdot) \)
  - Errors in estimating the underlying probability distribution.
  - Even if \( P(\cdot) \) is learned accurately, critical values are difficult to determine analytically from the distribution.
Our hypothesis testing framework

• Model-based approach with empirically-estimated sampling distribution:
Our hypothesis testing framework

- Model-based approach with empirically-estimated sampling distribution:
  - Learn $P(G)$ from single observed network using model.

$G_{obs}$

learn

model
Our hypothesis testing framework

- Model-based approach with empirically-estimated sampling distribution:
  - Learn $P(G)$ from single observed network using model.
  - \textbf{Sample a set of $N_s$ graphs from the learned model.}
Our hypothesis testing framework

• Model-based approach with empirically-estimated sampling distribution:
  – Learn $P(G)$ from single observed network using model.
  – Sample a set of $N_s$ graphs from the learned model.
  – **Calculate the likelihood of each network and sort them.**
Our hypothesis testing framework

- Model-based approach with empirically-estimated sampling distribution:
  - Learn $P(G)$ from single observed network using model.
  - Sample a set of $N_s$ graphs from the learned model.
  - Calculate the likelihood of each network and sort them.
Our hypothesis testing framework

- Model-based approach with empirically-estimated sampling distribution:
  - Learn $P(G)$ from single observed network using model.
  - Sample a set of $N_s$ graphs from the learned model.
  - Calculate the likelihood of each network and sort them.
  - Determine critical value from the empirical distribution.
Our hypothesis testing framework

- Model-based approach with empirically-estimated sampling distribution:
  - Learn $P(G)$ from single observed network using model.
  - Sample a set of $N_s$ graphs from the learned model.
  - Calculate the likelihood of each network and sort them.
  - Determine critical value from the empirical distribution.
  - Test hypotheses using the estimated critical value.

$G_{obs}$ → learn → model → $N_s$ → $\hat{P}_{obs}(G)$ → ordered $P_{obs}(G)$ → Reject $H_0$
Implementation

- We use the **mixed Kronecker Product Graph Model** (mKPGM; *Moreno et al. KDD‘13*)
  - mKPGMs are capable of learning *both* the mean and variance of an underlying graph population
  - It is also *efficient* to sample networks from mKPGMs, which facilitates estimation of empirical sampling distributions
Implementation

• We use the **mixed Kronecker Product Graph Model** (mKPGM; *Moreno et al. KDD’13*)
  
  – mKPGMs are capable of learning *both* the mean and variance of an underlying graph population
  
  – It is also *efficient* to sample networks from mKPGMs, which facilitates estimation of empirical sampling distributions
  
  – **However, mKPGM likelihoods are intractable to compute accurately, since they involve averaging over all possible graph permutations**—so we need an alternative likelihood to use as a test statistic
Implementation

- Our alternative likelihood calculates graph probability based on statistics of the graph (permutation invariant)

\[ X = \{ \overline{d}, \overline{cc}, \overline{g} \} \]
Implementation

- Our alternative likelihood calculates graph probability based on statistics of the graph (permutation invariant)

\[ X = \{ \overline{d}, \overline{cc}, \overline{g} \} \]

- **MBDE**: We approximate the empirical distribution of the graph statistics by a multivariate normal distribution

- **MBLE**: We approximate the empirical distribution of each graph statistic independently with a normal distribution.
Methodology

• We tested our approach on four different datasets, but here we only show results on an Email dataset.

• We evaluated our method against five baseline models and also compare mKPGMs against three alternative graph models.

• Evaluation measures:
  – False positive rate: % networks drawn from the null distribution that are incorrectly classified as being drawn from the alternative distribution.
  – Specificity: % networks drawn from the alternative distribution that are correctly classified.

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ True</th>
<th>$H_0$ False</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ Accepted</td>
<td>True Positive (TP)</td>
<td>False negative (FN)</td>
</tr>
<tr>
<td>$H_0$ Rejected</td>
<td>False positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td>FP rate</td>
<td>$\frac{FP}{(FP+TP)}$</td>
<td>Specificity=$\frac{TN}{(FN+TN)}$</td>
</tr>
</tbody>
</table>
Experimental results: Purdue email

- 189 email networks (1 per day) from 08/22/2011 to 02/28/2012
- Categorized the networks into *weekdays* and *weekends*.
- Learned models from a randomly selected weekday network.
- Use hypothesis tests to classify remaining networks as weekday or weekend (based on learned model).
Experimental results: Purdue email

- 189 email networks (1 per day) from 08/22/2011 to 02/28/2012
- Categorized the networks into *weekdays* and *weekends*.
- Learned models from a randomly selected weekday network.
- Use hypothesis tests to classify remaining networks as weekday or weekend (based on learned model).
Experimental results: Purdue email

• 189 email networks (1 per day) from 08/22/2011 to 02/28/2012
• Categorized the networks into *weekdays* and *weekends*.
• Learned models from a randomly selected weekday network.
• Use hypothesis tests to classify remaining networks as weekday or weekend (based on learned model).
Experimental results: Purdue email

- 189 email networks (1 per day) from 08/22/2011 to 02/28/2012
- Categorized the networks into weekdays and weekends.
- Learned models from a randomly selected weekday network.
- Use hypothesis tests to classify remaining networks as weekday or weekend (based on learned model).

<table>
<thead>
<tr>
<th>Model</th>
<th>false positive</th>
<th>specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test weekdays</td>
<td>test weekends</td>
</tr>
<tr>
<td>$mKPGM_{MBDE}$</td>
<td>12%</td>
<td>77%</td>
</tr>
<tr>
<td>$mKPGM_{MBLE}$</td>
<td>14%</td>
<td>77%</td>
</tr>
<tr>
<td>$DD$</td>
<td>68%</td>
<td>94%</td>
</tr>
<tr>
<td>$PNC$</td>
<td>89%</td>
<td>100%</td>
</tr>
<tr>
<td>$KPGM_{MBDE}$</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>$KPGM_{MBLE}$</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>$ERGM_{MBDE}$</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>$ERGM_{MBLE}$</td>
<td>99%</td>
<td>100%</td>
</tr>
<tr>
<td>$CL_{MBDE}$</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>$CL_{MBLE}$</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Most models rejected the null hypothesis for almost all networks which results in a high false positive rate.
Analysis of learned models

KPGM and CL models are biased

ERGM models did not learn the variance
Analysis of learned models

mKPGM models learn the distribution accurately. Why the errors?
Incorrect rejected weekdays: academic breaks, holidays, and special occasion days (Valentine’s day).
Analysis of learned models

Incorrectly accepted weekends: academic deadlines (weekends after final grades) and beginning of semesters (weekends before).
Conclusions

• We developed a novel, model-based hypothesis testing framework to determine if a new network is sampled from the same underlying graph distribution as a previously observed network.

• The results confirm that $\text{mKPGM}_{\text{MBDE}}$ and $\text{mKPGM}_{\text{MBLE}}$ learn the underlying graph distribution along with its respective variance, from a single network observation.

• This means the statistical hypothesis tests can correctly classify networks that are drawn from the same network population, while rejecting networks that are generated from alternative distributions.
Thanks for your attention

Email questions to: smorenoa@cs.purdue.edu

Code available at www.cs.purdue.edu/homes/smorenoa/codes.html