Using Bayesian Network Representation for Effective Sampling from Generative Network Models
Pablo Robles-Granda, Sebastian Moreno, Jennifer Neville, CS Department, Purdue University

Introduction
- Generative Network Models (GNMs) allow to sample random networks.
  - $E_i$ represents the existence of an edge $e_i$ between nodes $V'_i \in V$ and $V'_j \in V'$, where $P(E_i) = \pi_i$
  - Sampling – realization of the RVs $E_i$ for all $i$
  - Some GNMs are simple. However, GNMs that truly capture real-world network characteristics are more complex (complex interaction of random variables, RVs,).
- Bayesian Networks (BNs) could be used to represent the RVs’ interaction.
- Benefits: Insights derived from BN research for inference/sampling:
  - Compact representation of RVs relations
  - Take advantage of independence relations for sampling
  - Possible disadvantages:
    - BN inference is NP-hard. Not all GNM problems could be solved in reasonable time
  - Our analysis:
    - Mapping of GNM to BN
    - Using CSI property is not efficient enough for sampling
    - Deterministic CSD reduces the sampling space. This allows for faster sampling

Background:
- Bayesian Networks: Let $X_1, X_2, \ldots, X_n$ be a topological ordering of the nodes in the BN. Then, $X_i$ is independent of $(X_1, X_2, \ldots, X_{i-1},\text{pa}(X_i))|\text{pa}(X_i)$.
- Independence Properties:
  - Conditional Independence CI: Implicit to the BN
  - Context Specific Independence CSI: Let $X$, $Y$ and $W$ be distinct sets of RVs. Then $X \perp Y|W = w$ (which reads: $X$ is context-specific independent of $Y$ given $W = w$) if $P(X|Y, W = w) = P(X|W = w)$ whenever $P(Y, W = w) > 0$.
- GNM sampling:
  - A GNM samples network $G$ from the net-distribution $P(G)$.
  - Examples: Erdos-Renyi, Chung-Lu, BTER, mKPGM

Our approach: representing GNM as BN
- Ground level: Modeled by $P^e$, this level has $|V'|^2$ variables $Z_{ij}^{(0)}$
- All other layers $\lambda$ contain $(\lambda + 1)^2$ variables each
- The parameter of the BN is fully determined by the GNM parameters

Analysis of naïve, CI, CSI, DCSD sampling

NAÏVE SAMPLING
- Generate all CPT tables to derive the matrix of $N^2$ RVs at the last layer (edge probabilities).
- Impractical: exponential growth in the number of RVs per layer.

SAMPLING USING CI
- Sample can be done in the topological ordering of the network.
- RVs with the same parent are independent. The order in which these particular subset is sampled is not important (could be changed).
- The number of RVs increase at each layer of the hierarchy.
- Number or RVs to be sampled: $|V|^2 \cdot k^\ell$

SAMPLING USING CSI
- At each iteration of the sampling process the following CSI relation defines a context for mKPGM sampling:
  - $P(Z_{ij}^{(m)}|Z_{ij}^{(m-1)}, \ldots, Z_{ij}^{(0)}, \text{pa}(Z_{ij}^{(m)})) = 1 \forall Z_{ij}^{(m)}$
- Generates considerable the size of the CPT in a naïve sampling process.
- At the last layer ($\lambda = K - \ell$) the number of RVs is linear in the size of hierarchy: $K - \ell + 1$ (in comparison, naïve sampling, has $2^\ell$).

SAMPLING USING DCSD
- Condition: $P(Z_{ij}^{(m)}|Z_{ij}^{(m-1)} = 0, j, \lambda) = P(Z_{ij}^{(m)}|Z_{ij}^{(m-1)} = 0, \lambda)$ where $P(Z_{ij}^{(m)}|Z_{ij}^{(m-1)} = 1, j, \lambda) > 0 \forall j, \lambda$
- Sample each layer iteratively.
- Randomly sample an RV context per layer in the hierarchy.
- RVs with the same parent are independent. The order in which these particular subset is sampled is not important (could be changed).
- Sample can be done in the topological ordering of the network.
- Realization of the variables in a BN through DCSD with three levels corresponding to three tying iterations of mKPGM model.
- In the figure:
  - Realization or the variables in a BN through DCSD with three levels corresponding to three tying iterations of mKPGM model.
  - RVs that are set to 0 cause their descendants in the BN to be set to 0

Complexity analysis

<table>
<thead>
<tr>
<th>Property</th>
<th>Number of RVs</th>
<th>parent combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>$2^\ell$</td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td>CSI</td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td>DCSD</td>
<td>$\sum_{\lambda} N^2_{ij} \cdot k^\ell$</td>
<td>1</td>
</tr>
<tr>
<td>Random</td>
<td>$\sum_{\lambda} N^2_{ij} \cdot k^\ell$</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion, Current and Future Work
- Showed GNM-to-BN reduction
- Sampling can be done using naïve, CI, CSI, and DCSD approaches
- Although sampling with CSI is better than naïve, verifying CSI condition is also expensive. (CSD is simply the complementary concept of CSI)
- Using DCSD leads to a more efficient sampling method (e.g. mKPGM,)
- Correct sampling guaranteed by incorporating group probability sampling
- Open questions:
  - Some problems, other than sampling, in DCSD may be hard (subject of analysis in our ongoing work)
  - How dependencies in the network sampling may affect network characteristics
  - How problems in network sampling with BN-characteristics relate to other areas of BN research