Online Supplement: Degree Distribution Preserving Network Sampling: The Case of Relational Learning

Appendix A: Formulation for the Illustrative Example: NS_{eg}

$\mathcal{V}$: $\{1,2,3,4,5,6,7,8,9,10\}$
$\mathcal{E}$: $\{(1,2),(1,3),(1,4),(1,5),(1,6),(1,7),(1,9),(1,10),(2,3),(2,6),(2,9),(3,4),(3,5),$
$(3,6),(3,9),(4,5),(4,6),(4,7),(4,8),(4,9),(5,6),(5,7),(6,9),(7,9)\}$
N : 10
n : 5
$\mathcal{D}$: $\{d_1,d_2,d_3,d_4\}$
n_k : $n_1=1, n_2=1.5, n_3=2, n_4=0.5$
p_k : $p_1=\frac{2}{10}, p_2=\frac{3}{10}, p_4=\frac{4}{10}, p_4=\frac{1}{10}$
l_k : $l_1=0, l_2=0.25, l_3=0.50, l_4=0.75$
u_k : $u_1=0.2499, u_2=0.4999, u_3=0.7499, u_4=0.9999$

Table A.1 Illustrative Example: Parameters

x_i : $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$
$y_{i,j}$: $y_{1,2},y_{1,3},y_{1,4},y_{1,5},y_{1,6},y_{1,7},y_{1,9},y_{1,10},y_{2,3},y_{2,6},y_{2,9},y_{3,4},y_{3,5},$
$y_{3,6},y_{3,9},y_{4,5},y_{4,6},y_{4,7},y_{4,8},y_{4,9},y_{5,6},y_{5,7},y_{6,9},y_{7,9}$
$z_{i,k}$: $z_{1,1},z_{1,2},z_{1,3},z_{1,4},z_{2,1},z_{2,2},z_{2,3},z_{2,4},$
$z_{3,1},z_{3,2},z_{3,3},z_{3,4},z_{4,1},z_{4,2},z_{4,3},z_{4,4},$
$z_{5,1},z_{5,2},z_{5,3},z_{5,4},z_{6,1},z_{6,2},z_{6,3},z_{6,4},$
$z_{7,1},z_{7,2},z_{7,3},z_{7,4},z_{8,1},z_{8,2},z_{8,3},z_{8,4},$
$z_{9,1},z_{9,2},z_{9,3},z_{9,4},z_{10,1},z_{10,2},z_{10,3},z_{10,4}$

Table A.2 Illustrative Example: Decision Variables

min: Δ
subject to

constraint set 1

$$x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10} = 5 \quad (1.1)$$

constraint sets 2(i), 2(ii), 2(iii)

$x_1 \geq y_{1,2}$ (2.i.1)	$x_2 \geq y_{1,2}$ (2.ii.1)	$x_1+x_2-1 \leq y_{1,2}$ (2.iii.1)
$x_1 \geq y_{1,3}$ (2.i.2)	$x_3 \geq y_{1,3}$ (2.ii.2)	$x_1+x_3-1 \leq y_{1,3}$ (2.iii.2)
$x_1 \geq y_{1,4}$ (2.i.3)	$x_4 \geq y_{1,4}$ (2.ii.3)	$x_1+x_4-1 \leq y_{1,4}$ (2.iii.3)
$x_1 \geq y_{1,5}$ (2.i.4)	$x_5 \geq y_{1,5}$ (2.ii.4)	$x_1+x_5-1 \leq y_{1,5}$ (2.iii.4)
$x_1 \geq y_{1,6}$ (2.i.5)	$x_6 \geq y_{1,6}$ (2.ii.5)	$x_1+x_6-1 \leq y_{1,6}$ (2.iii.5)
$x_4 \geq y_{1,7}$ (2.i.6)	$x_7 \geq y_{1,7}$ (2.ii.6)	$x_1+x_7-1 \leq y_{1,7}$ (2.iii.6)
$x_1 \geq y_{1,9}$ (2.i.7)	$x_9 \geq y_{1,9}$ (2.ii.7)	$x_1+x_9-1 \leq y_{1,9}$ (2.iii.7)
$x_1 \geq y_{1,10}$ (2.i.8)	$x_{10} \geq y_{1,10}$ (2.ii.8)	$x_1+x_{10}-1 \leq y_{1,10}$ (2.iii.8)
$x_2 \geq y_{2,3}$ (2.i.9)	$x_3 \geq y_{2,3}$ (2.ii.9)	$x_2+x_3-1 \leq y_{2,3}$ (2.iii.9)
$x_2 \geq y_{2,6}$ (2.i.10)	$x_6 \geq y_{2,6}$ (2.ii.10)	$x_2+x_6-1 \leq y_{2,6}$ (2.iii.10)
$x_2 \geq y_{2,9}$ (2.i.11)	$x_9 \geq y_{2,9}$ (2.ii.11)	$x_2+x_9-1 \leq y_{2,9}$ (2.iii.11)
$x_3 \geq y_{3,4}$ (2.i.12)	$x_4 \geq y_{3,4}$ (2.ii.12)	$x_3+x_4-1 \leq y_{3,4}$ (2.iii.12)
$x_3 \geq y_{3,4}$ (2.i.13)	$x_5 \geq y_{3,5}$ (2.ii.13)	$x_3+x_5-1 \leq y_{3,5}$ (2.iii.13)
$x_3 \geq y_{3,6}$ (2.i.14)	$x_6 \geq y_{3,6}$ (2.ii.14)	$x_3+x_6-1 \leq y_{3,6}$ (2.iii.14)
$x_3 \geq y_{3,9}$ (2.i.15)	$x_9 \geq y_{3,9}$ (2.ii.15)	$x_3+x_9-1 \leq y_{3,9}$ (2.iii.15)
$x_4 \geq y_{4,5}$ (2.i.16)	$x_5 \geq y_{4,5}$ (2.ii.16)	$x_4+x_5-1 \leq y_{4,5}$ (2.iii.16)
$x_4 \geq y_{4,6}$ (2.i.17)	$x_6 \geq y_{4,6}$ (2.ii.17)	$x_4+x_6-1 \leq y_{4,6}$ (2.iii.17)
$x_4 \geq y_{4,7}$ (2.i.18)	$x_7 \geq y_{4,7}$ (2.ii.18)	$x_4+x_7-1 \leq y_{4,7}$ (2.iii.18)
$x_4 \geq y_{4,8}$ (2.i.19)	$x_8 \geq y_{4,8}$ (2.ii.19)	$x_4+x_8-1 \leq y_{4,8}$ (2.iii.19)
$x_4 \geq y_{4,9}$ (2.i.20)	$x_9 \geq y_{4,9}$ (2.ii.20)	$x_4+x_9-1 \leq y_{4,9}$ (2.iii.20)
$x_5 \geq y_{5,6}$ (2.i.21)	$x_6 \geq y_{5,6}$ (2.ii.21)	$x_5+x_6-1 \leq y_{5,6}$ (2.iii.21)
$x_5 \geq y_{5,7}$ (2.i.22)	$x_7 \geq y_{5,7}$ (2.ii.22)	$x_5+x_7-1 \leq y_{5,7}$ (2.iii.22)
$x_6 \geq y_{6,9}$ (2.i.23)	$x_9 \geq y_{6,9}$ (2.ii.23)	$x_6+x_9-1 \leq y_{6,9}$ (2.iii.23)
$x_7 \geq y_{7,9}$ (2.i.24)	$x_9 \geq y_{7,9}$ (2.ii.24)	$x_7+x_9-1 \leq y_{7,9}$ (2.iii.24)

constraint set 3

$z_{1,1}+z_{1,2}+z_{1,3}+z_{1,4} = x_1$ (3.1)	$z_{2,1}+z_{2,2}+z_{2,3}+z_{2,4} = x_2$ (3.2)
$z_{3,1}+z_{3,2}+z_{3,3}+z_{3,4} = x_3$ (3.3)	$z_{4,1}+z_{4,2}+z_{4,3}+z_{4,4} = x_4$ (3.4)
$z_{5,1}+z_{5,2}+z_{5,3}+z_{5,4} = x_5$ (3.5)	$z_{6,1}+z_{6,2}+z_{6,3}+z_{6,4} = x_6$ (3.6)
$z_{7,1}+z_{7,2}+z_{7,3}+z_{7,4} = x_7$ (3.7)	$z_{8,1}+z_{8,2}+z_{8,3}+z_{8,4} = x_8$ (3.8)
$z_{9,1}+z_{9,2}+z_{9,3}+z_{9,4} = x_9$ (3.9)	$z_{10,1}+z_{10,2}+z_{10,3}+z_{10,4} = x_{10}$ (3.10)

constraint set 4

$$y_{1,2}+y_{1,3}+y_{1,4}+y_{1,5}+y_{1,6}+y_{1,7}+y_{1,9}+y_{1,10} \geq 5(0.00z_{1,1}+0.25z_{1,2}+0.50z_{1,3}+0.75z_{1,4}) \quad (4.1)$$
$$y_{1,2}+y_{2,3}+y_{2,6}+y_{2,9} \geq 5(0.00z_{2,1}+0.25z_{2,2}+0.50z_{2,3}+0.75z_{2,4}) \quad (4.2)$$
$$y_{1,3}+y_{2,3}+y_{3,4}+y_{3,5}+y_{3,6}+y_{3,9} \geq 5(0.00z_{3,1}+0.25z_{3,2}+0.50z_{3,3}+0.75z_{3,4}) \quad (4.3)$$
$$y_{1,4}+y_{3,4}+y_{4,5}+y_{4,6}+y_{4,7}+y_{4,8}+y_{4,9} \geq 5(0.00z_{4,1}+0.25z_{4,2}+0.50z_{4,3}+0.75z_{4,4}) \quad (4.4)$$
$$y_{1,5}+y_{3,5}+y_{4,5}+y_{5,6}+y_{5,7} \geq 5(0.00z_{5,1}+0.25z_{5,2}+0.50z_{5,3}+0.75z_{5,4}) \quad (4.5)$$
$$y_{1,6}+y_{2,6}+y_{3,6}+y_{4,6}+y_{5,6}+y_{6,9} \geq 5(0.00z_{6,1}+0.25z_{6,2}+0.50z_{6,3}+0.75z_{6,4}) \quad (4.6)$$
$$y_{1,7}+y_{4,7}+y_{5,7}+y_{7,9} \geq 5(0.00z_{7,1}+0.25z_{7,2}+0.50z_{7,3}+0.75z_{7,4}) \quad (4.7)$$
$$y_{4,8} \geq 5(0.00z_{8,1}+0.25z_{8,2}+0.50z_{8,3}+0.75z_{8,4}) \quad (4.i.8)$$
$$y_{1,9}+y_{2,9}+y_{3,9}+y_{4,9}+y_{6,9}+y_{7,9} \geq 5(0.00z_{9,1}+0.25z_{9,2}+0.50z_{9,3}+0.75z_{9,4}) \quad (4.9)$$
$$y_{1,10} \geq 5(0.00z_{10,1}+0.25z_{10,2}+0.50z_{10,3}+0.75z_{10,4}) \quad (4.10)$$

constraint set 5

$$y_{1,2}+y_{1,3}+y_{1,4}+y_{1,5}+y_{1,6}+y_{1,7}+y_{1,9}+y_{1,10} \leq 5(0.2499z_{1,1}+0.4999z_{1,2}+0.7499z_{1,3}+0.9999z_{1,4}) \quad (5.1)$$
$$y_{1,2}+y_{2,3}+y_{2,6}+y_{2,9} \leq 5(0.2499z_{2,1}+0.4999z_{2,2}+0.7499z_{2,3}+0.9999z_{2,4}) \quad (5.2)$$
$$y_{1,3}+y_{2,3}+y_{3,4}+y_{3,5}+y_{3,6}+y_{3,9} \leq 5(0.2499z_{3,1}+0.4999z_{3,2}+0.7499z_{3,3}+0.9999z_{3,4}) \quad (5.3)$$
$$y_{1,4}+y_{3,4}+y_{4,5}+y_{4,6}+y_{4,7}+y_{4,8}+y_{4,9} \leq 5(0.2499z_{4,1}+0.4999z_{4,2}+0.7499z_{4,3}+0.9999z_{4,4}) \quad (5.4)$$
$$y_{1,5}+y_{3,5}+y_{4,5}+y_{5,6}+y_{5,7} \leq 5(0.2499z_{5,1}+0.4999z_{5,2}+0.7499z_{5,3}+0.9999z_{5,4}) \quad (5.5)$$
$$y_{1,6}+y_{2,6}+y_{3,6}+y_{4,6}+y_{5,6}+y_{6,9} \leq 5(0.2499z_{6,1}+0.4999z_{6,2}+0.7499z_{6,3}+0.9999z_{6,4}) \quad (5.6)$$
$$y_{1,7}+y_{4,7}+y_{5,7}+y_{7,9} \leq 5(0.2499z_{7,1}+0.4999z_{7,2}+0.7499z_{7,3}+0.9999z_{7,4}) \quad (5.7)$$
$$y_{4,8} \leq 5(0.2499z_{8,1}+0.4999z_{8,2}+0.7499z_{8,3}+0.9999z_{8,4}) \quad (5.8)$$
$$y_{1,9}+y_{2,9}+y_{3,9}+y_{4,9}+y_{6,9}+y_{7,9} \leq 5(0.2499z_{9,1}+0.4999z_{9,2}+0.7499z_{9,3}+0.9999z_{9,4}) \quad (5.9)$$
$$y_{1,10} \leq 5(0.2499z_{10,1}+0.4999z_{10,2}+0.7499z_{10,3}+0.9999z_{10,4}) \quad (5.10)$$

constraint set 6

$$\Delta \geq \tfrac{1}{5}\left(z_{1,1}+\ldots+z_{10,1}-n_1\right) \quad (6.1)$$
$$\Delta \geq \tfrac{1}{5}\left(z_{1,1}+\ldots+z_{10,1}+z_{1,2}+\ldots+z_{10,2}-n_1-n_2\right) \quad (6.2)$$
$$\Delta \geq \tfrac{1}{5}\left(z_{1,1}+\ldots+z_{10,1}+z_{1,2}+\ldots+z_{10,2}+z_{1,3}+\ldots+z_{10,3}-n_1-n_2-n_3\right) \quad (6.3)$$
$$\Delta \geq \tfrac{1}{5}\left(z_{1,1}+\ldots+z_{10,1}+z_{1,2}+\ldots+z_{10,2}+z_{1,3}+\ldots+z_{10,3}+z_{1,4}+\ldots+z_{10,4}-n_1-n_2-n_3-n_4\right) \quad (6.4)$$

constraint set 7

$$\Delta \geq \tfrac{1}{5}\left(n_1-z_{1,1}-\ldots-z_{10,1}\right) \quad (7.1)$$
$$\Delta \geq \tfrac{1}{5}\left(n_1+n_2-z_{1,1}-\ldots-z_{10,1}-z_{1,2}-\ldots-z_{10,2}\right) \quad (7.2)$$
$$\Delta \geq \tfrac{1}{5}\left(n_1+n_2+n_3-z_{1,1}-\ldots-z_{10,1}-z_{1,2}-\ldots-z_{10,2}-z_{1,3}-\ldots-z_{10,3}\right) \quad (7.3)$$
$$\Delta \geq \tfrac{1}{5}\left(n_1+n_2+n_3+n_4-z_{1,1}-\ldots-z_{10,1}-z_{1,2}-\ldots-z_{10,2}-z_{1,3}-\ldots-z_{10,3}-z_{1,4}-\ldots-z_{10,4}\right) \quad (7.4)$$

In addition, we need Δ to be non-negative and binary requirements on the x_i, $y_{i,j}$, and $z_{i,k}$ variables.

Appendix B: Proofs of Propositions and Theorems

B.1. Proof of Proposition 1

PROPOSITION 1. *The removal of v will (a) decrease the relative degrees of all neighbors $w \in \mathcal{A}(v)$, and (b) increase the relative degrees of all non-neighbors $w' \notin \mathcal{A}(v)$.*

Proof: (a) When node v is removed from a network with N nodes, the degrees of all its neighbors decrease by 1. Consider a neighbor w of v with degree δ; w has relative degree $\delta_{r_w} = \frac{\delta}{N}$ in the original network. The removal of v reduces the degree of w to $(\delta_{r_w}N - 1)$, and makes the new relative degree $\left(\frac{\delta_{r_w}N-1}{N-1}\right)$, which is less than δ_{r_w}.

(b) When node v is removed, the degrees of all non-neighbors are unchanged. That is, given a non-neighbor of v, $w' \notin \mathcal{A}(v)$ with degree δ, the removal of v changes the relative degree from $\frac{\delta}{N}$ to $\left(\frac{\delta}{N-1}\right)$ (which is greater than $\frac{\delta}{N}$). □

B.2. Proof of Proposition 2

PROPOSITION 2. *If* $\frac{|\mathcal{A}(w)|-1}{N-1} \geq l_{k(w)}\ \forall w \in \mathcal{A}(v)$ *and* $\frac{|\mathcal{A}(w')|}{N-1} < u_{k(w')}\ \forall w' \in \mathcal{V} \setminus \mathcal{A}(v)$*, the removal of* v *will not affect* Δ*.*

Proof: From Proposition 1(a), we know that the relative degree of the neighbors of node v will decrease when v is removed; the new relative degree of $w \in \mathcal{A}(v)$ is $\frac{|\mathcal{A}(w)|-1}{N-1}$. If this new value does not fall below the lower bound of bin $k(w)$ (i.e., if $\frac{|\mathcal{A}(w)|-1}{N-1} \geq l_{k(w)}$), then node w will remain in bin $k(w)$. Similarly, Proposition 1(b) says that the relative degree of the non-neighbors w' of node v will increase. If the new relative degree does not exceed the upper bound of the bin to which w' belongs (i.e., if $\frac{|\mathcal{A}(w')|}{N-1} < u_{k(w')}$), then node w' will remain in bin $k(w')$. Therefore, if the first condition holds for all neighbors of v and the second holds for all its non-neighbors, all remaining nodes will stay in their original bins when v is removed, leaving the value of Δ unchanged. □

B.3. Proof of Proposition 3

PROPOSITION 3. *3 The value of* Δ *will not change when deleting* $\mathcal{H}_k$ *if (i)* $u_{k(w)} > \frac{|\mathcal{A}(w)|-\mu_w}{N-(N_k-n_k)} \geq l_{k(w)}\ \forall v \in \mathcal{H}_k, w \in \mathcal{A}(v)$*, and* $w \notin \mathcal{H}_k$*, and (ii)* $u_{k(w')} > \frac{|\mathcal{A}(w')|}{N-(N_k-n_k)}\ \forall v \in \mathcal{H}_k, w' \notin \mathcal{A}(v)$*, and* $w' \notin \mathcal{H}_k$*.*

Proof: Consider a node $v \in \mathcal{H}_k$. The removal of v reduces the relative degrees of all nodes in $\mathcal{A}(v)$ (Proposition 1). Consider a node w that is a neighbor to one or more nodes in $\mathcal{H}_k$. The new relative degree of w is $\frac{|\mathcal{A}(w)|-\mu_w}{N-(N_k-n_k)}$. As $\mu_w < (N_k - n_k)$, the new relative degree can be lower or higher than that before the removal of the nodes in $\mathcal{H}_k$. Therefore, the removal of these nodes has to ensure that the new relative degree of every such w stays between $l_{k(w)}$ and $u_{k(w)}$ – i.e., $u_{k(w)} > \frac{|\mathcal{A}(w)|-\mu_w}{N-(N_k-n_k)} \geq l_{k(w)}$ for all nodes that have at least one neighbor in $\mathcal{H}_k$. The new relative degree of a node w' that is not a neighbor to any node in $\mathcal{H}_k$ is $\frac{|\mathcal{A}(w')|}{N-(N_k-n_k)}$, and this has to remain below $u_{k(w')}$ for all such nodes w' to ensure that none of them move to a bin to the right. If both these conditions hold, none of the remaining nodes will move to a different bin when the nodes in $\mathcal{H}_k$ are removed from $\mathcal{G}$, and the value of Δ will not be affected. □

B.4. Proof of Theorem 1

THEOREM 1. *If* OPTSAMPLE *finds a solution, it is optimal and the associated KS distance is 0.*

Proof: Since KS distance is non-negative, the lowest possible value is zero. If the conditions in Proposition 3 can be applied successfully to identify the set $\mathcal{H}_k$ of nodes to be removed for every bin k, the exact number of nodes that need to be removed from bin k is removed, without impacting any other bin. As a result, relative degree bin k of the sample stays identical to relative degree bin

k of the original. After processing all K bins, we would have removed $(N - n)$ nodes, leaving a sample of size n whose relative degree distribution is identical to that of the original, ensuring that the KS distance between the two distributions is zero. □

Appendix C: Illustrative Example for OptSample

In this section, we illustrate how to extract a sample of 6 nodes from the 9-node example network in Figure C.1. We represent the original histogram using 3 bins, with $[0, 0.435)$, $[0.435, 0.76)$, and $[0.76, 1.00)$ as the bin boundaries in this illustration. Each original bin has 3 nodes, with nodes 1, 2, and 3 in bin 1, nodes 4, 5, and 6 in bin 2, and nodes 7, 8, and 9 in bin 3. If a perfect sample exists, it has to have 2 nodes in each bin. We will show how to extract a perfect sample by deleting one node from each bin, starting with bin 1 and ending with bin 3.

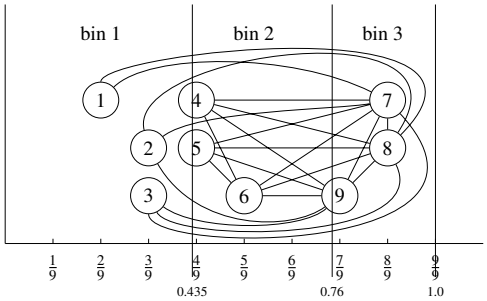
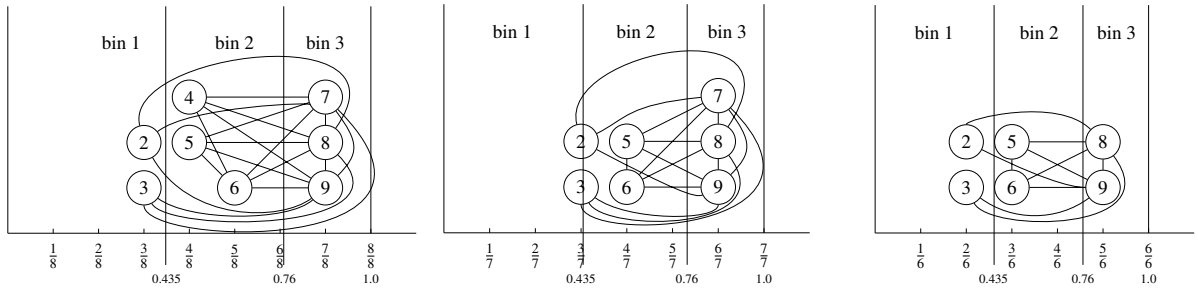


Figure C.1: Illustrative Example: OptSample **(Original Network)**

Step 1 (bin 1): Node 1 has two neighbors (7 and 8), while nodes 2 and 3 have three (7, 8, 9). The removal of node 1 will result in nodes 7 and 8 having updated relative degrees of $\frac{7}{8} > 0.76$. Nodes 2–6 and 9 will have increased relative degrees: $\frac{3}{8}$ for nodes 2 and 3, $\frac{4}{8}$ for nodes 4 and 5, $\frac{5}{8}$ for node 6, and $\frac{7}{8}$ for node 9. However, none of these nodes will move to a different bin if node 1 is removed. If either node 2 or node 3 is removed on the other hand, the relative degree of node 9 drops to $\frac{6}{8} = 0.75 < 0.76$, making it shift left, to bin 2. Therefore, $\mathcal{H}_1 = \{1\}$ is the best node to delete from bin 1 according to Proposition 3. The resulting network (with 8 nodes) is as shown in Figure C.2a.

Step 2 (bin 2): An analysis similar to that in Step 1 shows that removing either node 4 or node 5 will not make their neighbors change bins. For example, removing node 4 changes the relative degrees of nodes 7, 8, and 9 to $\frac{6}{7} = 0.85 > 0.76$, and that of node 6 to $\frac{4}{7} = 0.57 > 0.435$ (as these are its neighbors). Non-neighbor nodes 2 and 3 will have relative degrees of $\frac{3}{7} = 0.42 < 0.435$, while node 5 will have a relative degree of $\frac{4}{7} = 0.57 < 0.76$. Removing node 6 on the other hand changes the relative degrees of nodes 4 and 5 to $\frac{3}{7} = 0.42 < 0.435$, making them shift to bin 1. Consequently, setting $\mathcal{H}_2 = \{4\}$ and removing node 4 from bin 2 keeps the KS distance unchanged at zero. The network after Step 2 has 7 nodes, and is shown in Figure C.2b.

Step 3 (bin 3): We can remove any of the three nodes in bin 3 without affecting the bin membership of the other nodes. Suppose we choose to remove node 7, i.e., $\mathcal{H}_3 = \{7\}$. Figure C.2c shows the final sample network of 6 nodes. As we have managed to arrive at exactly the number of nodes required in each bin of the sample, the relative degree distribution of the sample is identical to that of the original network, and the KS distance is 0.



(a) Network After Step 1 **(b) Network After Step 2** **(c) Network After Step 3**

Figure C.2 Illustrative Example: OPTSAMPLE **(Additional Steps)**

Appendix D: Heuristics

Algorithm 2: BATCHDELETION (BD)

input : The original or intermediate network $\mathcal{G}$ and associated histogram, sample size n
output The sample network
:
while *Size of sample network* $> n$ **do**
  for *Bin* $k \in \{1, \ldots, K\}$ **do**
    Calculate $\zeta_k = \min\left\{\sum_{i=1}^{k}(n'_i - n_i), n'_i, \sum_{i=1}^{K}(n'_i) - n\right\}$
    if $\zeta_k \leq 0$ **then**
      Move to bin $k+1$ (i.e., to Step 2).
    if $\zeta_k = n'_k$ **then**
      Delete all n_k′ nodes from bin k and move to bin $k+1$ (i.e., to Step 2).
    if $\zeta_k < n'_k$ **then**
      Determine $\mathcal{S}_{avg}$ scores of all n'_k nodes in bin k.
      Sort the nodes in decreasing values of the scores.
      Delete the ζ_k nodes with the highest $\mathcal{S}_avg$ values.
    Update the histogram for the remaining nodes in the sample network.

D.1. Computational Complexity of BD

In each pass, the following computations would be required for a bin. The computational complexities associated with checking the size of a bin or the network, calculating ζ_k, and deleting a node, are negligible. The worst-case complexity of computing $\mathcal{S}(v)$ in Step 2 for all the nodes in a bin is $O(N^2)$ because we must iterate through each neighbor of every node in the bin. The complexity of sorting the nodes based on their $\mathcal{S}(v)$ scores in Step 2 is $O(Nlog(N))$. Lastly, the complexity of updating the histogram after the requisite number of nodes are deleted from the bin in Step 2 is $O(NK)$. Thus, the overall complexity associated with node deletions for a bin is $O(N^2 + Nlog(N) + NK) \sim O(N^2)$ since $K << N$. As we have to iterate over K bins in each pass, and as there could be at most $N - n$ passes required, the overall worst-case complexity is $O(KN^3)$.

D.2. Computational Complexity of MBD

In practice, the complexity is close to that of BATCHDELETION depending on the value of m. This is because for each bin, recomputing the $\mathcal{S}_{avg}$ scores, sorting the remaining nodes in the bin, and

Algorithm 3: MiniBatchDeletion (MBD)

input : The original or intermediate network $\mathcal{G}$ and associated histogram, sample size n
output The sample network
:
Define m and m_f.
while *Size of sample network* $> n$ **do**
  for *Bin* $k \in \{1, \ldots, K\}$ **do**
    Calculate $\zeta_k = \min\{\sum_{i=1}^{k}(n'_i - n_i), n'_i, \sum_{i=1}^{K}(n'_i) - n\}$
    if $\zeta_k \leq 0$ **then**
      Move to bin $k+1$ (i.e., to Step 3).
    Calculate $\zeta_{k,m} = \lfloor \frac{\zeta_k}{m} \rfloor$.
    for *i from* 1 *to* $m-1$ **do**
      Compute the $\mathcal{S}_{avg}$ values for all remaining nodes in the bin.
      Sort the nodes in decreasing order of $\mathcal{S}_{avg}$.
      Delete the $\zeta_{k,m}$ nodes with highest $\mathcal{S}_{avg}$ value.
      Update the histogram for the remaining sample network.
      If the revised $n'_k < \zeta_{k,m}$ then delete all n'_k nodes and move to bin $k+1$ (i.e., to Step 3).
    Calculate $\zeta_{k,m_f} = \lfloor \frac{\zeta_{k,m}}{m_f} \rfloor$.
    for *i from* 1 *to* m_f **do**
      Compute the $\mathcal{S}_{avg}$ values for all remaining nodes in the bin.
      Sort the nodes in decreasing order of $\mathcal{S}_{avg}$.
      Delete the ζ_{k,m_f} nodes with highest $\mathcal{S}_{avg}$ value.
      Update the histogram for the remaining sample network.
      If the revised $n'_k < \zeta_{k,m_f}$ then delete all n'_k nodes and move to bin $k+1$ (i.e., to Step 3).

revising the histogram, are conducted at most $m - 1 + m_f$ times in MBD. Thus, the worst-case complexity is $O(N.K.(m-1+m_f)(N^2 + NlogN + K.N)) \sim O((m+m_f)KN^3)$. In practice, we find both heuristics are very efficient, as demonstrated by our experiments.

Appendix E: Impact of Sample Size on KS Distances and Sample Selection Times

We conducted experiments on three different datasets to explore trends in KS distances and solution times for sampling. The first set of experiments involved the moderately sized dataset MD. We extracted samples of sizes 10% to 90% in increments of 10% using heuristic MBD. The results of these experiments are shown in Table E.2a. These results show that time increases with sample size. The effect of sample size on KS distance is less obvious particularly for smaller sample sizes, though KS distance is observed to decrease steadily as the sample sizes increase beyond a point (for samples larger than 40%). The results on the larger datasets (Table E.2b) are less conclusive along both dimensions as the changes are not monotonic. We find that KS distances remain low even for small sample sizes, while the solution times remain quite stable in all our experiments. This is encouraging – as samples are typically needed when networks are large, these results suggest that extracting small samples is effective from the perspectives of both solution time and KS distance.

Dataset	Sample Size	Benchmarks					Heuristics			
		FS	RW	SB	ESi	RN	IP	LP	BD	MBD
EC	10%	0	0	0	0	0	14,417	5	1	0
	20%	0	0	0	0	0	14,218	14	1	0
	30%	0	0	0	0	0	14,402	18	1	0
MD	10%	0	0	3	4	0	14,490	199	11	17
	20%	1	1	5	7	1	14,788	959	12	4
	30%	3	1	10	10	1	14,403	5,422	16	12
MFB	10%	1	0	6	11	1	14,413	1,482	7	8
	20%	5	1	13	24	1	14,400	3,902	8	9
	30%	12	1	21	45	2	14,403	216	8	36
GDFB	10%	9	6	245	68	8	-	-	49	52
	20%	19	6	564	161	10	-	-	55	61
	30%	29	8	840	289	13	-	-	58	253
GDHR	10%	6	1	69	66	6	14,410	14,410	36	39
	20%	13	3	155	151	10	-	-	44	47
	30%	23	4	249	262	13	-	-	45	210
GC	10%	81	73	6,748	1,492	58	-	-	1,080	923
	20%	134	97	14,109	4,491	106	-	-	1,312	1,079
	30%	207	165	20,507	72,52	149	-	-	1,499	4,594
OM	0.18%	2,758	3,559	-	-	7,291	-	-	6,554	11,590
	0.20%	4,455	3,384	-	-	7,798	-	-	6,824	9,849
	0.25%	4,975	3,991	-	-	9,085	-	-	10,903	11,363

Table E.1 Sample Extraction Times (Seconds)

Sample Size	KS Distance	Time (Secs)
10%	0.0380	11
20%	0.0162	12
30%	0.0181	15
40%	0.0452	21
50%	0.0237	25
60%	0.0177	26
70%	0.0117	27
80%	0.0099	30
90%	0.0039	32

(a) Dataset MD

GC			OM		
Sample Size	KS Distance	Time (Secs)	Sample Size	KS Distance	Time (Secs)
10,762 (10%)	0.0091	3,272	10,941 (0.00100%)	0.0056	18,392
16,143 (15%)	0.0119	4,705	13,097 (0.00125%)	0.0043	19,678
21,523 (20%)	0.0085	4,157	15,279 (0.00150%)	0.0045	20,546
26,904 (25%)	0.0159	7,479	17,462 (0.000020%)	0.0027	20,257
32,285 (30%)	0.0066	1,957	19,645 (0.00125%)	0.0013	19,960
37,665 (35%)	0.0220	8,759	21,828 (0.00250%)	0.0078	21,110

(b) Datasets GC and OM

Table E.2 KS Distances and Sample Extraction Times

Appendix F: Variances of Degrees in Samples

Table F.1 presents the variances of the distributions associated with the results presented in Table 4. As noted in the paper, the deviations of the variances relative to the original – i.e., $\frac{\left|\sigma^2_{heuristic}-\sigma^2_{original}\right|}{\sigma^2_{original}}$ – over the 13 experiments where all approaches provided solutions were lowest for BD and MBD, at 0.62 and 0.65 respectively. RN performed the best among the benchmarks, with a value of 1.0. All other approaches were substantially worse, with values of 16.13 (FS), 24.79 (RW), 11.88 (SB), 19.31 (ESi), and 8.91 (LP). The significantly lower deviations associated with BD and MBD suggest that our approaches preserve the shape of the original distribution better than the benchmarks.

Appendix G: Comparisons With Other Distance Measures

Table G.1 presents the impact of the heuristics and benchmarks on four common measures of inter-distribution distance: Kullback-Leibler divergence, Wasserstein distance, Jensen-Shannon diver-

Dataset	Original	Sample Size	Benchmarks					IP	Heuristics		
			FS	RW	SB	ESi	RN		LP	BD	MBD
EC	1.35×10^{-03}	10%	1.22×10^{-2}	1.45×10^{-2}	1.21×10^{-2}	1.61×10^{-2}	3.77×10^{-6}	1.72×10^{-3}	1.11×10^{-3}	3.10×10^{-3}	2.42×10^{-3}
		20%	8.32×10^{-3}	1.00×10^{-2}	8.02×10^{-3}	1.09×10^{-2}	2.03×10^{-6}	1.18×10^{-3}	3.71×10^{-4}	1.85×10^{-3}	9.33×10^{-4}
		30%	7.05×10^{-3}	8.28×10^{-3}	7.04×10^{-3}	7.92×10^{-3}	0.00	1.27×10^{-3}	1.43×10^{-3}	2.32×10^{-3}	1.72×10^{-3}
MD	7.97×10^{-05}	10%	1.95×10^{-3}	2.19×10^{-3}	1.89×10^{-3}	2.29×10^{-3}	2.33×10^{-9}	4.17×10^{-6}	5.03×10^{-5}	5.74×10^{-5}	2.53×10^{-5}
		20%	8.81×10^{-4}	9.56×10^{-4}	8.93×10^{-4}	1.08×10^{-3}	2.03×10^{-9}	3.55×10^{-5}	2.04×10^{-5}	3.28×10^{-5}	3.58×10^{-5}
		30%	5.06×10^{-4}	6.11×10^{-4}	5.78×10^{-4}	6.20×10^{-4}	0.00	8.90×10^{-6}	1.93×10^{-5}	2.09×10^{-5}	1.80×10^{-5}
MFB	1.38×10^{-06}	10%	1.00×10^{-4}	1.99×10^{-4}	3.89×10^{-5}	1.10×10^{-4}	0.00	1.05×10^{-6}	1.10×10^{-4}	8.77×10^{-7}	2.36×10^{-6}
		20%	2.88×10^{-5}	6.23×10^{-5}	2.94×10^{-5}	4.25×10^{-5}	1.54×10^{-10}	6.15×10^{-7}	2.45×10^{-6}	6.70×10^{-7}	9.17×10^{-7}
		30%	1.74×10^{-5}	2.10×10^{-5}	1.55×10^{-5}	2.20×10^{-5}	0.00	9.63×10^{-7}	2.60×10^{-7}	1.03×10^{-6}	1.31×10^{-6}
GDFB	2.23×10^{-06}	10%	4.58×10^{-5}	4.89×10^{-5}	3.65×10^{-5}	5.06×10^{-5}	4.65×10^{-11}	0.00	1.64×10^{-6}	4.63×10^{-6}	6.73×10^{-6}
		20%	2.72×10^{-5}	2.82×10^{-5}	2.48×10^{-5}	3.10×10^{-5}	2.51×10^{-11}	0.00	9.58×10^{-6}	1.27×10^{-6}	1.43×10^{-6}
		30%	1.65×10^{-5}	1.93×10^{-5}	1.68×10^{-5}	1.92×10^{-5}	0.00	1.29×10^{-9}	0.00	1.98×10^{-6}	1.85×10^{-6}
GDHR	1.08×10^{-07}	10%	1.55×10^{-6}	1.74×10^{-6}	1.02×10^{-6}	1.77×10^{-6}	1.23×10^{-11}	7.73×10^{-8}	3.15×10^{-6}	2.48×10^{-7}	2.68×10^{-7}
		20%	9.91×10^{-7}	1.16×10^{-6}	8.15×10^{-7}	1.14×10^{-6}	9.22×10^{-12}	-	-	8.42×10^{-8}	8.08×10^{-8}
		30%	6.78×10^{-7}	7.97×10^{-7}	5.98×10^{-7}	7.68×10^{-7}	0.00	-	-	8.58×10^{-8}	9.55×10^{-8}
GC	2.92×10^{-05}	10%	2.95×10^{-3}	3.02×10^{-3}	1.56×10^{-3}	3.42×10^{-3}	3.83×10^{-11}	-	-	3.47×10^{-5}	2.30×10^{-5}
		20%	1.11×10^{-3}	1.05×10^{-3}	9.87×10^{-4}	1.30×10^{-3}	8.59×10^{-12}	-	-	1.94×10^{-5}	1.54×10^{-5}
		30%	5.18×10^{-4}	5.91×10^{-4}	5.05×10^{-4}	6.01×10^{-4}	0.00	-	-	2.71×10^{-5}	1.92×10^{-5}
OM	7.57×10^{-09}	10%	1.18×10^{-4}	1.10×10^{-4}	-	-	0.00	-	-	3.30×10^{-7}	1.67×10^{-6}
		20%	1.17×10^{-4}	1.49×10^{-4}	-	-	0.00	-	-	1.28×10^{-6}	2.88×10^{-6}
		30%	9.15×10^{-5}	8.68×10^{-5}	-	-	0.00	-	-	4.82×10^{-7}	1.62×10^{-6}

Table F.1 Variances of Relative Degrees of Samples from Table 4

gence, and Jeffreys divergence. These results show that the heuristic for minimizing KS distance also provides samples with lower distances based on these metrics, compared to the benchmarks.

Distance Measure	Sample Size	MBD	BD	FS	RW	SB	ESi
Kullback-Leibler	10,762 (10%)	0.0618	**0.0180**	5.0963	5.0662	2.4945	5.6515
Wasserstein		0.0030	**0.0012**	0.0163	0.0171	0.0137	0.0171
Jensen-Shannon		0.1196	**0.0680**	0.5994	0.6249	0.5083	0.6239
Jeffreys		0.0720	**0.0445**	3.3543	3.4864	1.7770	3.7358
Kullback-Leibler	21,523 (20%)	0.0146	**0.0091**	2.0145	1.9676	1.7014	2.4020
Wasserstein		0.0012	**0.0007**	0.0140	0.0141	0.0130	0.0146
Jensen-Shannon		0.0629	**0.0516**	0.5176	0.5187	0.4780	0.5422
Jeffreys		0.0379	**0.0295**	1.5773	1.5604	1.3140	1.8424
Kullback-Leibler	32,285 (30%)	**0.0081**	0.0082	1.1240	1.2479	1.0253	1.2754
Wasserstein		**0.0007**	**0.0007**	0.0125	0.0129	0.0116	0.0129
Jensen-Shannon		**0.0450**	0.0465	0.4427	0.4587	0.4130	0.4606
Jeffreys		**0.0134**	0.0157	0.9630	1.0594	0.8495	1.0774

Table G.1 Impact on Other Distance Measures (dataset GC)

Appendix H: Experiments on Synthetic Networks

We conducted experiments on networks generated through models commonly used for generating synthetic networks (Newman 2016); specifically, we generate networks using the following models – (i) Erdős-Rényi (ER), (ii) Power cluster law (PC), (iii) Watts-Strogatz (WS), and (iv) Bárabasi-Albert (BA). The results of these experiments are provided in Table H.1.

We note that all of these models have some recognized drawbacks vis-à-vis real networks (Lim et al. 2015). For example, Erdős-Rényi networks do not have heavy tails, while the Watts-Strogatz model is known to produce networks with small-world properties but with unrealistic degree distributions. Neither of these models generate networks with power-law degree distributions which is common in real-world networks. The Bárabasi-Albert model was proposed to overcome the

Model	Dataset	Sample Size	MBD	FS	RW	SB	ESi
Erdős-Rényi	G1	10,000	0.3632	0.3619	0.3412	0.3573	0.3019
		20,000	0.1592	0.2514	0.2351	0.2474	0.2286
		30,000	0.1237	0.1952	0.1974	0.1888	0.1780
	G2	10,000	0.2937	0.3101	0.3197	0.3062	0.2815
		20,000	0.1894	0.2229	0.2201	0.2231	0.2032
		30,000	0.1368	0.1697	0.1682	0.1645	0.1634
Power Cluster Law	G1	10,000	0.0141	0.4080	0.4064	0.3858	0.3322
		20,000	0.0036	0.1994	0.1840	0.1938	0.1783
		30,000	0.0008	0.1193	0.1104	0.1128	0.1090
	G2	10,000	0.0734	0.8083	0.7939	0.7966	0.7869
		20,000	0.0926	0.5344	0.4938	0.5151	0.5105
		30,000	0.0547	0.3510	0.3197	0.3339	0.3289
Watts-Strogatz	G1	10,000	0.4084	0.9227	0.9103	0.8030	0.6066
		20,000	0.4516	0.7777	0.7361	0.7429	0.5460
		30,000	0.3015	0.6672	0.6301	0.6535	0.4334
	G2	10,000	0.9565	0.4058	0.3987	0.4008	0.3415
		20,000	0.9142	0.3318	0.4186	0.3394	0.3018
		30,000	0.6997	0.2590	0.3980	0.2690	0.2415
Bárabasi-Albert	G1	10,000	0.0224	0.5574	0.5670	0.5592	0.4722
		20,000	0.0049	0.2822	0.2716	0.2810	0.2550
		30,000	0.0049	0.1806	0.1732	0.1821	0.1689
	G2	10,000	0.0423	0.6584	0.6812	0.6631	0.6516
		20,000	0.1092	0.5323	0.5369	0.5272	0.5182
		30,000	0.0907	0.4333	0.4294	0.4326	0.4188

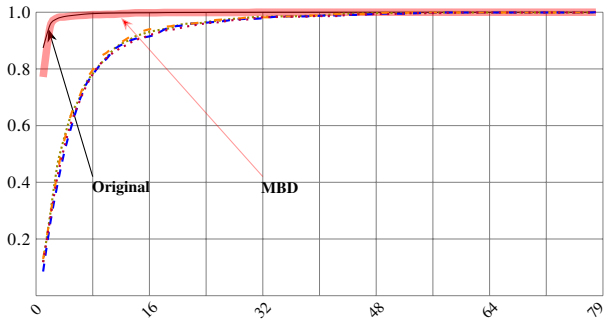
Table H.1 KS Distances: Synthetic Networks

deficiencies of these models, but it is known to not produce the high levels of clustering sometimes observed in real networks. We created eight synthetic networks (two for each model) with 100,000 nodes each, by varying parameters that control network density. We extracted 10%, 20%, and 30% samples from each network through MBD and the benchmarks. Except for one synthetic network, we find that the samples obtained using the heuristics we propose have KS distances considerably lower than the samples derived from the benchmarks. This was not the case for samples drawn from one of the two Watts-Strogatz networks – as noted earlier, we had developed our heuristic with right-skewed real-world networks in mind, and Watts-Strogatz networks are known to have unrealistic degree distributions (Lim et al. 2015). The average KS value across all 24 experiments is 0.230 for MBD compared to 0.36 or higher for the benchmarks. The corresponding values were 0.099 and 0.338 if we consider only the 18 experiments involving the Erdős-Rényi, the Power-Cluster law, and the Bárabasi-Albert networks. With one exception, MBD performs better on the sparser networks.

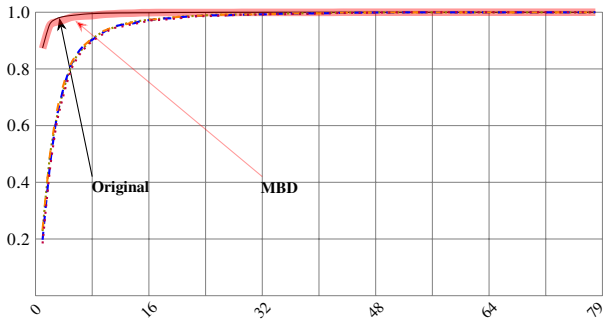
Appendix I: Cumulative Distributions of Original and Samples: Datasets MD and GC

We expand on Figure 7 in this appendix. Figures I.1 and I.2 present the original degree distributions for datasets MD and GC respectively, along with those of 10%, 20%, and 30% samples drawn using MBD and four benchmarks (FS, RW, SB, and ESi). The observation made with Figure 7 continues

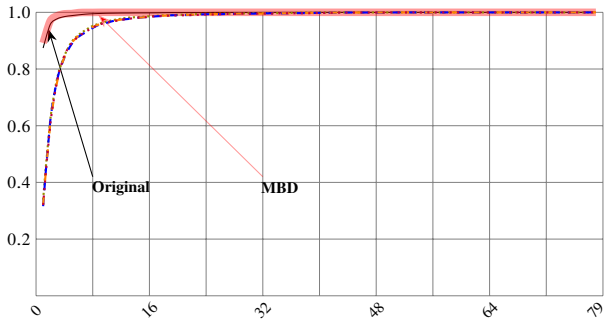
to hold – i.e., the distribution of the sample drawn using MBD is almost identical to that of the original network, while the others are not.



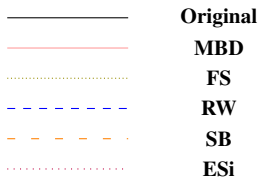
(a) Sample Size: 10%



(b) Sample Size: 20%

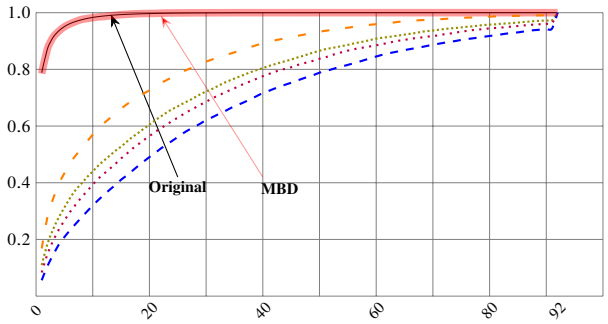


(c) Sample Size: 30%

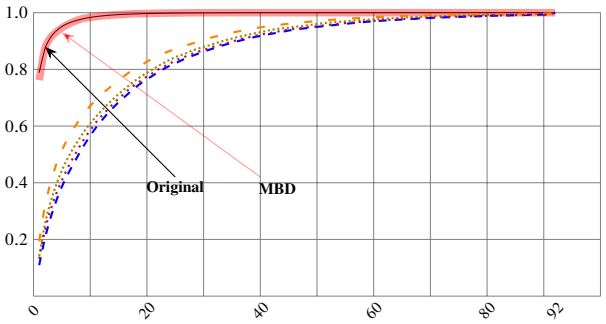


(d) Legend

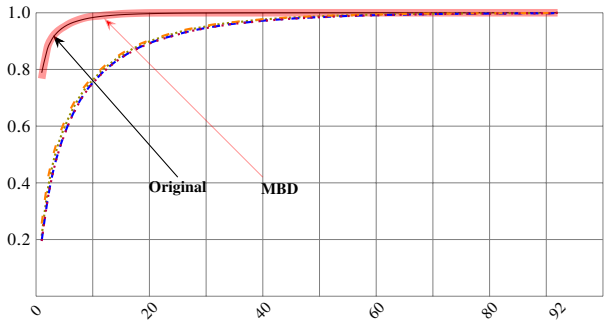
Figure I.1 Cumulative Probability vs Number of Bins (Dataset MD)



(a) Sample Size: 10%



(b) Sample Size: 20%



(c) Sample Size: 30%

(d) Legend

Figure I.2 Cumulative Probability vs Number of Bins (Dataset GC)