IX. Appendix

Theorem IV.1. *Log-likelihood F in Equation (17) is non-monotone submodular for all SEIRS-type models except SIS.*

Proof:

F has three types of terms; higher order terms from $\log(\mathcal{L}_e^j)$, quadratic or linear terms from $\log(\mathcal{L}_s^j)$ depending on $\mathcal{M}$ and linear terms from $\log(\mathcal{L}_i^j)$ and $\log(\mathcal{L}_r^j)$. F is non-monotone since linear and quadratic terms are either positive or negative depending on $\mathcal{M}$, transition distribution parameters and the terms from $\log(\mathcal{L}_s^j)$ that model the probability of susceptible nodes not being infected/exposed.

F is submodular when $F(A+x) - F(A) \geq F(B+x) - F(B)$ for every $A \subset B$ and for every $x \in U \setminus (A \cup B)$. To prove submodularity of F, we prove the submodularity of each term in F since summation of submodular functions is also submodular. Linear terms of F are unimodular, so they are submodular. Quadratic terms show up in $\log(\mathcal{L}_s^j)$ when $\mathcal{M}$ is loopy and when the model is not SIS, each quadratic term is one of the following: $Q(r_{v,j-1}, r_{u,j-1}) = \log(1-p_{uv})(1-r_{v,j-1})(1-r_{u,j-1})$, $Q(r_{v,j-1}, e_{u,j-1}) = \log(1-p_{uv})(1-r_{v,j-1})(1-e_{u,j-1})$ or $Q(r_{v,j-1}, i_{u,j-1}) = \log(1-p_{uv})(1-r_{v,j-1})(1-i_{u,j-1})$. All those terms are submodular since they satisfy the inequality $Q(0,0)+Q(1,1) \leq Q(0,1)+Q(1,0)$.

Then, we need to prove the submodularity of the higher-order terms that depend on G to prove submodularity of F. Higher-order terms appear in either $\log(\mathcal{L}_e^j)$ for bipartite models or $\log(\mathcal{L}_i^j)$ for non-bipartite diffusion models. Depending on $\mathcal{M}$, we need to prove either $T = s_{v,j-1} \log\left(1.0 - \mathcal{L}_{s2e}^{v,j,I}\ \mathcal{L}_{s2e}^{v,j,R}\right)$ or $T = s_{v,j-1} \log\left(1.0 - \mathcal{L}_{s2i}^{v,j,I}\ \mathcal{L}_{s2i}^{v,j,R}\right)$. Each variable might appear at two positions of T; either inside or outside the logarithm. When $\mathcal{M}$ is bipartite, each variable can only appear in one of those positions whereas it can appear in both positions for non-bipartite $\mathcal{M}$. Let $V_e = \bigcup_{u \in P(v) \cap I_j} e_{u,j-1}$, $V_r = \bigcup_{u \in P(v) \cap E_j} r_{u,j-1}$, x be the variable to be added, X be the current set of added variables, $K = \Pi_{V_e \cup V_r}(1-p_{uv})$ and $P_t = (1-p_{tv})^t$ for every $t \in V_e \cup V_r$, T is submodular as proven below.

- If x is outside the logarithm, let $A = \{a, b\}$ and $B = \{a, b, c\}$. Then, $T(A+x) = \log\left(1 - \frac{K}{P_a P_b P_x}\right)$, $T(B+x) = \log\left(1 - \frac{K}{P_a P_b P_c P_x}\right)$ and $T(A+x) - T(A) \geq T(B+x) - T(B)$ will be satisfied since $T(A+x) \geq T(B+x)$ and $T(A) = T(B) = 0$.
- If x is inside the logarithm, when $s_{v,j-1} \notin X$, submodularity is trivially satisfied since $T(A) = T(A+x) = T(B) = T(B+x) = 0$. When $s_{v,j-1} \in X$, let $A = \{a\}$ and $B = \{a, c\}$ ($A \subset B$), submodularity is satisfied as shown in Equation (39)–(41).

$$T(A+x) - T(A) \geq T(B+x) - T(B) \tag{39}$$

$$\log\left(\frac{1-\frac{K}{P_a P_x}}{1-\frac{K}{P_a}}\right) \geq \log\left(\frac{1-\frac{K}{P_a P_b P_x}}{1-\frac{K}{P_a P_c}}\right) \tag{40}$$

$$K P_a P_b (1-P_b)(1-P_a) \geq 0 \tag{41}$$

Then, F is submodular since each summation term including the higher-order ones is submodular.

■

Theorem IV.2. *History reconstruction from log-likelihood for SIS model can be expressed as submodular maximization under both packing and partition matroid constraints.*

Proof:

Quadratic terms $Q(s_{v,j-1}, s_{u,j-1})$ from $\mathcal{L}_s^j$ are supermodular for SIS but they can be turned into submodular ones as follows: We define new varible $i_{v,j-1}$ for every node $v \in \{S_j \cup I_j\}$ to represent whether v is infected at time $j-1$. Then, we obtain the new objective function F^* by replacing each supermodular $Q(s_{v,j-1}, s_{u,j-1}) = log(1-p_{uv}) s_{v,j-1}(1-s_{u,j-1})$ with $Q^*(s_{v,j-1}, s_{u,j-1}) = log(1-p_{uv}) s_{v,j-1} i_{v,j-1}$. We also add assignment constraints of $s_{v,j-1} + i_{v,j-1} = 1$ for every node $v \in \{S_j \cup I_j\}$ to make sure node v is either infected or susceptible at $j-1$. Each $Q^*(s_{v,j-1}, s_{u,j-1})$ in F^* is submodular since it satisfies the inequality $Q^*(0,0) + Q^*(1,1) \leq Q^*(0,1) + Q^*(1,0)$. Then, F^* is submodular since the rest of the higher-order terms are submodular as proven in Theorem IV.1. Assignment constraints define partition matroid and the problem of reconstructing history at time $j-1$ becomes submodular maximization under both partition matroid and existing packing constraints for SIS model.

■

Theorem IV.3. *Algorithm 1 has approximation guarantee of $k + \frac{S_0}{O}(1-k)$ for $k = \frac{1}{3}$ in terms of minimization of supermodular $-F$ for each of its iteration.*

Proof:

Let X be the set of elements returned by the non-monotone submodular maximization algorithm and $F(X) = -M$. We are interested in upper-bounding the supermodular minimization ratio ($\frac{M}{O}$) for $-F$. Since F_n is obtained by adding S_0 to each set in F, $\frac{F_n(X)}{F_n(X_{opt})} = \frac{S_0 - M}{S_0 - O} \geq k$ and we obtain $\frac{M}{O} <= k + \frac{S_0}{O}(1-k)$. Here, $\frac{S_0}{O}(1-k)$ makes the approximation ratio data-dependent and this ratio is the best we can achieve when k is tight for non-monotone submodular maximization. This data-dependent bound is also the best we can achieve in terms of supermodular minimization perspective since non-negative supermodular minimization problem cannot be approximated in constant factor unless $P = NP$ [1].

■

Theorem IV.4. $\log(\mathcal{L}_{j,k}^{in})$ *in Equation 18 is non-monotone submodular for all SEIRS-type models.*

Proof:

We prove the submodularity of $\log(\mathcal{L}_{j,k}^{in})$ by proving the submodularity of each of its summation terms. $\log(P(X_{j+1}|D_j))$ estimates the most probable diffusion snapshot at $j+1$ given D_j. It is a forward estimate and if we use the same variable naming as in Section IV-A, it becomes a linear function of X_{j+1} and thus submodular.

$\log(P(D_k|X_{k-1}))$ is same as F (17) in Section IV-A and it is submodular as proven in Theorem IV.1.

Every $\log(P(X_{t+1}|X_t))$ involves the variables from both time steps t and $t+1$. Here, we do not know the exact node states at both time steps so we define all possible state variables for every node for both time steps ($s_{v,t}$, $e_{v,t}$, $i_{v,t}$, $r_{v,t}$, $s_{v,t+1}$, $e_{v,t+1}$, $i_{v,t+1}$, $r_{v,t+1}$, $\forall v \in V$). $\log(P(X_{t+1}|X_t))$ can be expressed as in Equation 42 where the likelihoods are defined as in Equation (43)–(46). Each term in $\log(P(X_{t+1}|X_t))$ is additive and log-likelihood terms of endogenous transitions are submodular since they are quadratic terms with negative coefficient. Log-Likelihood terms of exogenous transitions are also submodular by following the submodularity proof of the higher-order terms from Theorem IV.1.

$$\log(P(X_{t+1}|X_t)) = \log(\mathcal{L}_s^{t+1}) \ + \log(\mathcal{L}_e^{t+1}) \ + \log(\mathcal{L}_i^{t+1}) \\ \log(\mathcal{L}_r^{t+1}) \tag{42}$$

$$\mathcal{L}_{exo} = \prod_{u \in P(v)} (1 - p_{uv})^{i_{u,t} s_{v,t}}$$

$$\mathcal{L}_e^{t+1} = \prod_{v \in V} \left((1 - e2i_v)^{e_{v,t} e_{v,t+1}} \ (1.0 - \mathcal{L}_{exo})^{e_{v,t+1}}\right) \tag{43}$$

$$\mathcal{L}_s^{t+1} = \prod_{v \in V} \left(\mathcal{L}_{exo}^{s_{v,t+1}} (r2s_v)^{r_{v,t} s_{v,t+1}}\right) \tag{44}$$

$$\mathcal{L}_r^{t+1} = \prod_{v \in V} \left((i2r_v)^{i_{v,t} r_{v,t+1}} (1.0 - r2s_v)^{r_{v,t} r_{v,t+1}}\right) \tag{45}$$

$$\mathcal{L}_i^{t+1} = \prod_{v \in V} \left((e2i_v)^{e_{v,t} i_{v,t+1}} (1.0 - i2r_v)^{i_{v,t} i_{v,t+1}}\right) \tag{46}$$

■

Theorem V.1. *Prize Collecting Dominating Set Vertex Cover (PCDSVC) is NP-hard, and it can be approximated by* $O(\log(|V^*|))$.

Proof:

PCDSVC is NP-hard since its special case *Dominating Set* is NP-hard that is obtained when all edge weights are 0 ($w_{uv} = 0$).

Given PCDSVC problem over graph $G^* = (V^*, E^*)$, we construct *Minimum Hitting Set* instance (S, C) as follows: We define the set of elements as $S = \{v \in V^*\} \cup \{e \in E^*\}$ where the cost of each each item in E^* is w_u for every $u \in V^*$ and w_{uv} for every $(u,v) \in E^*$. Subsets $C = C_1 \cup C_2$ of S are defined as: $C_1 = \{e_u, e_v, e_{uv}\}$ $,\forall (u,v) \in E^*$ and $C_2 = \{e_u, u \in P(v) \cup \{v\}\}$ $,\forall v \in V^*$. This reduction is linear time, approximation preserving and the solution of this *Minimum Hitting Set* gives us the solution for PCDSVC. Here $|S| = |E^*| + |V^*|$ and **Greedy** method for *Set Cover* approximates this problem by $\log(|S|) + 1 \approx O(\log(|E^*| + |V^*|)) + 1 \approx O(\log(|V^*|)) + 1$.

One can also easily show that each *Minimum Hitting Set* instance can be reduced to PCDSVC and this reduction is also approximation preserving. Then, *Minimum Hitting Set* and PCDSVC are *equivalent* under linear reduction and this approximation ratio for PCDSVC is the best we can achieve unless P=NP [2].

■

Theorem V.2. *The Taylor expansion relaxation of* (17) *for bipartite diffusion models can be expressed as s-t mincut.*

Proof:

Minimization problem for bipartite $\mathcal{M}$ has objective F_{bi} as seen in Equation 47. F_{bi} is a regular function [3]: when expressed as the summation of first and second-order terms as in Equation 48, each second order term $E^{u,v}(s_{v,j-1}, i_{u,j-1})$ satisfies $E^{u,v}(0,0) + E^{u,v}(1,1) \leq E^{u,v}(0,1) + E^{u,v}(1,0)$ in regular functions. Regular functions can be solved optimally by transforming it into s-t mincut [3]. Transformation is as follows:

$$\min \ -F_{bi} = \sum_{(u,v) \in E^*} \frac{1}{\log(1 - p_{uv})} (1 - i_{u,j-1}) s_{v,j-1} \\ + \sum_{v \in E_j \cup S_j} w_v s_{v,j-1} + \sum_{v \in I_j \cup R_j} w_v i_{v,j-1} \tag{47}$$

$$-F_{bi} = \sum_{u \in I_j \cup R_j, v \in E_j \cup S_j} E^{u,v}(i_{u,j-1}, s_{v,j-1}) \ + \\ \sum_{v \in I_j \cup R_j} E^v(i_{v,j-1}) + \sum_{v \in S_j \cup E_j} E^v(s_{v,j-1}) \tag{48}$$

We define new directed graph $G' = (V', E')$ where $V' = V^* \cup \{s\} \cup \{t\}$. For every $v \in V^*$, we add edge (s,v) with weight $E^v(1)$ if $E^v(1) > 0$ and add edge (v,t) with weight $-E^v(1)$ if $-E^v(1) < 0$. For every $u \in I_j \cup R_j$ and $v \in S_j \cup E_j$, we add edge (u,v) with weight $E^{u,v}(0,1)$. s-t mincut solution of this graph gives us the resulting node partition; after the cut edges removed, variables of the nodes that are reachable from s are assigned 1 and the variables of the nodes that have a path to t are assigned 0.

■

References

[1] L. A. Wolsey and G. L. Nemhauser, *Integer and Combinatorial Optimization*. Wiley-Interscience, 1999.

[2] U. Feige, "A threshold of ln n for approximating set cover," *Journal of the ACM*, vol. 45, no. 4, pp. 634–652, Jul. 1998.

[3] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 65–81, 2004.