The Most Persistent Soft-Clique in a Set of Sampled Graphs

Novi Quadrianto[1] | Chao Chen[2] | Christoph Lampert[2]
1: Cambridge Machine Learning Group, University of Cambridge | 2: IST Austria (Institute of Science and Technology Austria)

UNIVERSITY OF CAMBRIDGE

Abstract

- We introduce the concept of most persistent soft-clique. This is subset of vertices, that 1) is almost fully or at least densely connected, 2) occurs in all or almost all graph instances, and 3) has the maximum weight;
- We present a measure of clique-ness, that essentially counts the number of edge missing to make a subset of vertices into a clique. With this measure, we show that the problem of finding the most persistent soft-clique problem can be cast either as: a) a max-min two person game optimization problem, or b) a min-min soft margin optimization problem;
- We show that both formulations lead to the same solution when using a partial Lagrangian method to solve the optimization problems;
- Our proposed approach has a direct application for searching characteristic subpatterns in a collection of potentially noisy graph data.

Motivating Example

Goal:

- To identify a clique of friends from, for example, video sequences, mobile-phone or location-based social network graph.

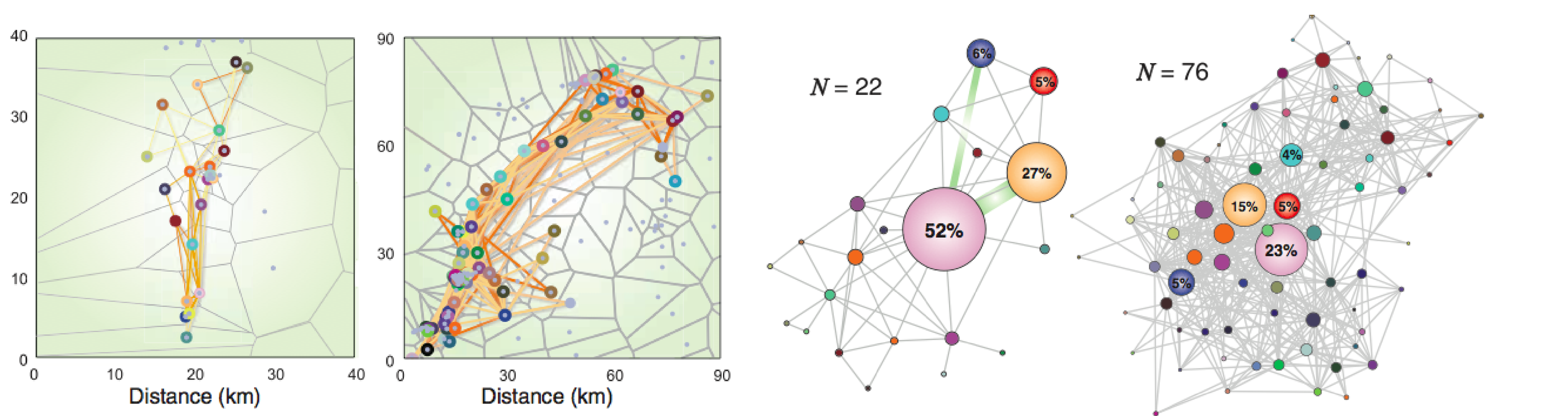
(Potential) Problems:

- Inconsistencies: different graph instances have different edge sets. For example, a person could have left the group temporarily due to other commitments, or the measurement itself could be faulty.

Example 1: Video Sequences Data

BIWI Walking Pedestrians dataset

Example 2: Mobile-phone Mobility Network Graph



Chaoming Song et al., Science 2010

Persistent Soft-Cliques

Notations

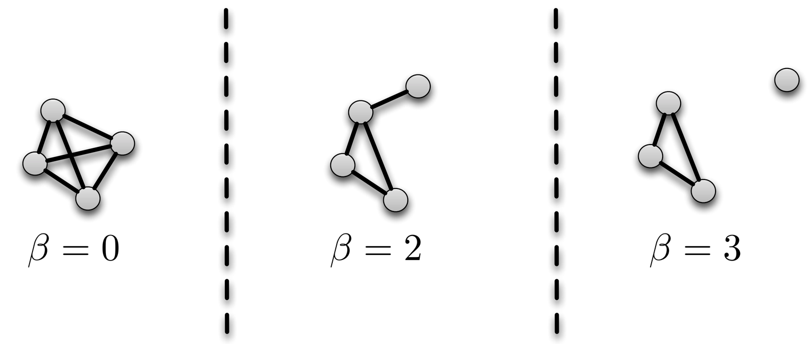
- For a set of vertices $\mathcal{V} = \{v_1, \ldots, v_n\}$, let $\mathcal{E}_t \subseteq \mathcal{V} \times \mathcal{V}$, for $t = 1, \ldots, T$, be multiple observed sets of edges;
- Let $k_t : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ be non-negative weight functions. We have $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t, k_t)$, for $t = 1, \ldots, T$;
- Let $S \subset \mathcal{V}$ be a vertex subset. Let $x_i \in \{0,1\}^{|\mathcal{V}|}$ be an indicator variable whether or not the vertex v_i is included as a clique, where $x_i = 1$ if $v_i \in S$, and $x_i = 0$ otherwise.

The problem of finding a maximum weighted clique in a weighted graph $(\mathcal{V}, \mathcal{E}, k)$ can be cast as the following optimization problem:

$$\max_{x \in \{0,1\}^{|\mathcal{V}|}} \sum_{1 \le i < j \le n} x_i x_j k(v_i, v_j) \;\text{ subject to }\; \sum_{1 \le i < j \le n} x_i x_j \mathbb{I}[(i,j) \notin \mathcal{E}] = 0 \quad (1)$$

Soft Clique-ness is a measure of how far a set of vertices is from being a clique, that is

$$\beta := \sum_{1 \le i < j \le n} x_i x_j \mathbb{I}[(i,j) \notin \mathcal{E}]$$



Persistency of a Clique over Time Given multiple instances of a graph, find a soft-clique that persists through time. We formalize this concept as either slack or two-person game, discussed in turn.

Slack Perspective We turn hard-cliques constraints in (1) into a soft-clique constraints by introducing *slack variables*, β_t, for $t = 1, \ldots, T$.

$$\min_{x \in \{0,1\}^{|\mathcal{V}|}} \min_{\beta \in \mathbb{R}_+^T} \underbrace{-\sum_{1 \le i < j \le n} x_i x_j k(v_i, v_j)}_{\text{Regularizer}} + \eta \underbrace{\|\beta\|_{L_p}^p}_{\text{Loss}}$$

$$\text{subject to } \sum_{1 \le i < j \le n} x_i x_j \mathbb{I}[(i,j) \notin \mathcal{E}_t] \le \beta_t \quad \forall t = 1, \ldots, T.$$

Two-Person Game Perspective In this perspective, two competing players: *inlier* and *outlier* are involved. The *inlier* player controls $x \in \{0,1\}^{|\mathcal{V}|}$ and aims at finding a group of variables with as large weight as possible. The *outlier* aims at reducing the objective value by controlling variables $\beta_1, \ldots, \beta_T$, which he or she can increase up a limit given by the number of edges missing to make x a clique.

$$\max_{x \in \{0,1\}^{|\mathcal{V}|}} \min_{\beta \in \mathbb{R}_+^T} \sum_{1 \le i < j \le n} x_i x_j k(v_i, v_j) - \sum_t \beta_t^p$$

$$\text{subject to } \sum_{1 \le i < j \le n} x_i x_j \mathbb{I}[(i,j) \notin \mathcal{E}_t] \ge \beta_t \quad \forall t = 1, \ldots, T.$$

Optimization

We replace the soft-cliqueness constraint by a Lagrangian. We then partially dualize the lower bound (upper bound) of slack (game) perspective by finding the stationary point with respect to only the primal variables β. We show the case of ℓ_2 measure ($p = 2$), for other cases, please refer to the paper.

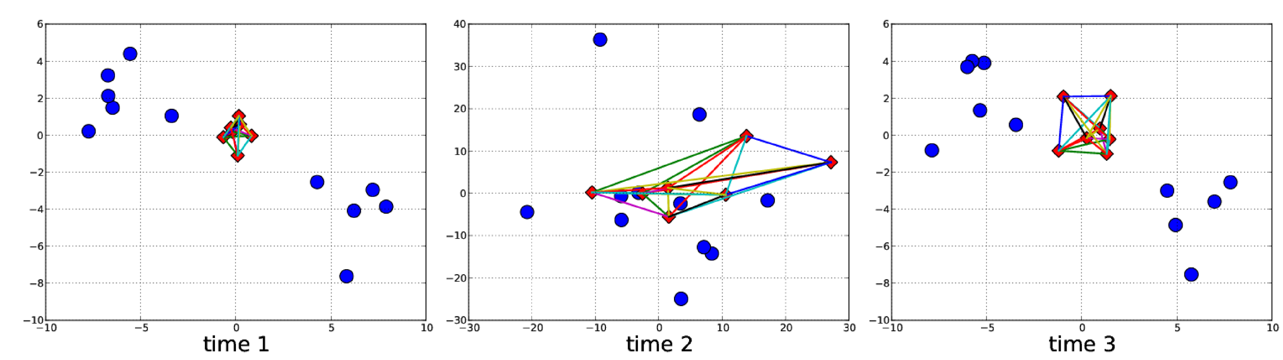
Algorithm: ℓ_2 Soft Clique-ness Measure

Input $\mathcal{G}_t(\mathcal{V}, \mathcal{E}_t, k_t)$ for $t = 1, \ldots, T$, N iterations, η constant
Compute the total similarity, $k(i,j) = \sum_t k_t(i,j)$
for $i = 1$ **to** N **do**
  Compute the measure, $c(i,j) = \sum_t \lambda_t \mathbb{I}[(i,j) \notin \mathcal{E}_t]$
  Solve $\operatorname{argmax}_x \{x^T K x - x^T C x\}$
  Update $\lambda_t \leftarrow 2\eta \sum_{ij} x_i x_j \mathbb{I}[(i,j) \notin \mathcal{E}_t]$
end for
Return $x \in \{0,1\}^{|\mathcal{V}|}$

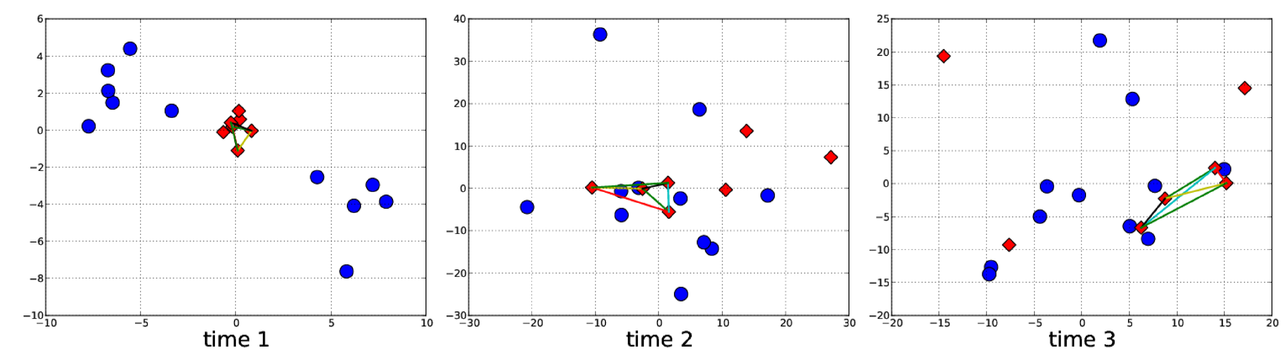
Experiments

Synthetic Data

High noise at time 2 and low noise at time 3



High noise for both time 2 and 3



Jaccard Index Metric

Data	GS	Soft ℓ_2
A	0.82±0.28	0.89±0.14
B	0.58±0.31	0.64±0.17
C	0.79±0.28	0.87±0.16
D	0.85±0.26	0.89±0.15

- Data: at time 1 are drawn from a Gaussian mixture with 3 components. At time 2 and 3, the data are corrupted with a random Gaussian noise.
- Baseline: graph shift algorithm, Liu et al. ICML 2010.

Real Social Network Data

- Data: a Brightkite location-based social network graph `http://snap.stanford.edu/data/loc-brightkite.html`.
- Results: We define different *after*-hours in a day as samples of the graphs. We represent a person with a bag of vectors, and use set kernels with sub-polynomial trick to reduce the diagonal dominance. We observe that our identified clique explains 23% of the friendship network that was collected based on the online public API, in comparison with 14% of a random null model.