A Novel Method for Vertex Clustering in Dynamic Networks

Devavrat Vivek Dabke
Princeton University

Olga Dorabiala
University of Washington

Introduction

We introduce spatiotemporal graph *k*-means (STG*k*M), a novel method for clustering vertices in a dynamic network. Drawing inspiration from classical *k*-means, we develop a technique to identify both short-term and long-lived communities that respect the overall dynamics of a graph.

Our method combines several desirable properties:

1. Temporal smoothness
2. Only one required parameter: *k*, the number of clusters
3. Multiscale analysis at the most granular and coarsest levels

Main Objectives

We introduce spatiotemporal graph *k*-means (STG*k*M), a novel method for clustering vertices in a dynamic network. Drawing inspiration from classical *k*-means, we develop a technique to identify both short-term and long-lived communities that respect the overall dynamics of a graph.

Our method combines several desirable properties:

1. Temporal smoothness
2. Only one required parameter: *k*, the number of clusters
3. Multiscale analysis at the most granular and coarsest levels

Spatiotemporal Graph *k*-means

Setup Given a dynamic graph we pre-compute the shortest journey $\delta^t(u, v)$ from vertex *u* to *v* starting at time *t* for all pairs of vertices and all time. We require *k* and optionally take two tuning parameters: λ, γ.

Phase 1 Solve optimization unified over space and time:

$$\min_{c,W} \sum_{u \in V} \sum_{j \in [k]} W^t_{u,j} \cdot \delta^t(u, c^t_j) \qquad \delta^{t-q}(c^{t-1}_j, c^t_j) \leq \lambda$$

$W \triangleq \{0,1\}^{|\mathbb{T}| \times |V| \times k}$ — All subsets of *V* of length *k* — $1 \leq q \leq \gamma$

Objective ***Constraint***

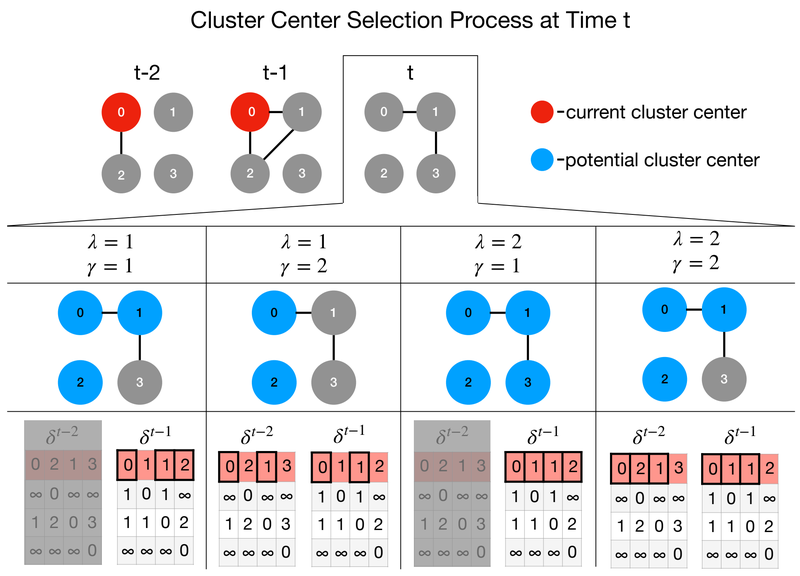

Figure 1. At time t, c_0^t is chosen based on c_0^{t-1} . The drift time window γ determines for how many previous time steps centers must be within maximum drift λ of one another. The objective function is evaluated for all potential cluster centers; the center that minimizes the objective is chosen.

Phase 2 We define a metric between vertices as the Hamming distance of their respective assignment histories *W*. With this metric, we perform agglomerative clustering to partition vertices into long-lived spatiotemporal clusters.

Theoretical Results

Theorem 1 Given a holding, non-stranding dynamic graph with *k* connected components, the partition induced by the optimal solution is exactly the connected components with sufficient time.

In this setting, vertices in different clusters eventually have infinite distance, so we can separate them.

Lemma 1 For two vertices u, v in distinct connected components, there exists some time step t_0 such that the distance between them is infinite.

If not true, these vertices would be connected by definition.

Lemma 2 In a self-connected dynamic network, connectivity* is an equivalence relation; the connected components are the respective equivalence classes.

Lemma 3 For two vertices in distinct connected components in a non-stranding, holding dynamic network, there exists a time step such that the distance between them is infinite after that timestep.

This result extends Lemma 1 to guarantee that vertices never become connected after becoming disconnected. It eliminates pathological cases, especially in very long-ranged networks.

Choosing *k*

STG*k*M is an unsupervised method but requires a known number of clusters. We use a technique analogous to the *Elbow Method* to automatically select an optimal value of *k,* which further obviates the need for manual intervention. We search for the global minimum objective value across sufficiently small *k*.

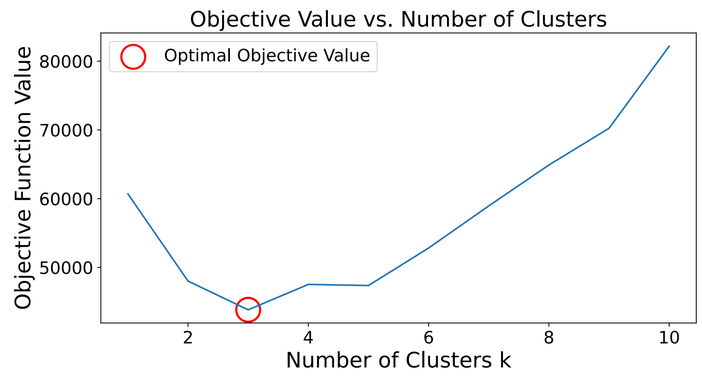

Figure 2. Objective function value versus number of clusters on a synthetic dataset.

Experimental Results

Synthetic Data We generate three cliques and then randomly perturb the graph. After running STG*k*M, we can identify the original cliques as they evolve over time. This example demonstrates the operation of this algorithm.

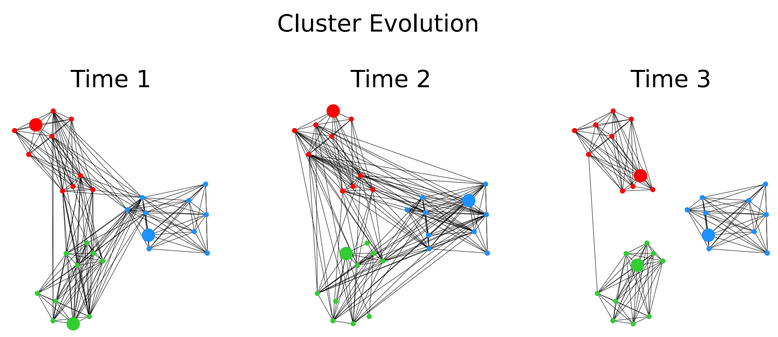

Figure 3. Three snapshots of a dynamic graph and the dynamic clustering as predicted by STGkM. Cluster centroids are identified by enlarged nodes.

Voting Data We form a graph of roll call votes from the United States House of Representatives. Each representative is a node, and nodes are connected if and only if they vote the same on a given issue. Each vote is a time step. We obtain three long-lived clusters: Republicans, Democrats, and a small band of "outlying" Democrats. We choose *k* using our method described above.

Figure 4. Similarity matrices of the short term clustering similarity between nodes in the Roll Call dataset. Rows and columns of the matrices are organized by detected long-term community membership.