Ilya Safro

Multiscale methods for large networks: response to epidemics and network generation problems

228 McAdams Hall
School of Computing
Clemson University
Clemson
SC 29634
isafro@clemson.edu
Alexander Gutfraind
Sven Leyffer
Lauren Meyers

Networks are a widely used type of abstraction for complex data. Optimization of different quantitative objectives on networks often plays a crucial role in network science, not only when a practical solution is needed, but also for a general understanding of structural and statistical features of networks. We present multiscale approaches for two problems: optimal response to epidemics, and network generation. Both approaches are inspired by AMG scheme reinforced by the algebraic distance connectivity strength.

Heuristic for optimal response to epidemics and cyber attacks. We consider a traditional infection-spread SIS model in which network nodes can be in one of two possible states, namely, *infected* and *susceptible*, and each node i is associated with probability of being infected at time t, $\phi_{i,t}$. This model has been extensively analyzed in epidemiology and adapted in the cyber security area for analysis of computer viruses propagation. We now formulate the optimization problem whose goal is to keep the level of infection at each node low while maximizing the (weighted) number of "alive" connections. This is motivated by the infection spread response policies that are often driven by the number of resources available for the realization. We denote the graph of a network by $G = (V, E, w)$, where V and E are sets of nodes and edgfes, respectively, and $w : V \to \mathbb{R}_{\geq 0}$ represents the strength of connectivity (such as the number of shared users in a cyber system). Assuming that the probabilities of infection transition from Γ_i (neighbors of i) to i are independent, the problem

is formulated as

$$\begin{aligned} \max_{x} \quad & \sum_{ij \in E} w_{ij} x_i x_j \\ \text{subject to} \quad & x_i - \prod_{j \in \Gamma_i} (1 - p_{ij} \phi_{j,t-1} x_j) \leq b_i \quad \forall i \in V, \\ & x_i \in \{0, 1\} \qquad\qquad\qquad\qquad \forall i \in V, \end{aligned} \tag{1}$$

where w_{ij} is the link weight between nodes i and j; b_i is a threshold for the level of allowed probability of infection at i; and x_i are binary variables (1 - if we decide to leave the node i functioning, 0 - closed, requiring special attention). In general, (1) is a nonconvex integer nonlinear program and known to be NP-complete. We demonstrate a strategy for designing fast multiscale methods for this class of problems. The refinement represents the collective improvement for sufficiently small subsets of variables. This phase can easily be performed in parallel by using the red-black order. Single subset refinement solves problem for subset of nodes by choosing a connected subgraph and fixing the *boundary conditions* for the rest of the nodes.

We evaluate our method on a set of small networks with known optimal solutions, two case studies (HIV spread and cyber infrastructure networks), and one massive data set. The two case study networks are typical complex network instances on which solving this particular optimization is of great practical importance. The massive dataset evaluation contains networks of different structures and sources that can potentially represent hard structures for the method. In all experiments our method improved best known results (quality and/or running time) significantly.

Multiscale Entropic Network Generation. High-quality, large-scale network data is often not available for scientists, because of economic, legal, technological, or other obstacles. For example, the human contact networks along with infectious diseases spread are notoriously difficult to estimate, and thus our understanding of the dynamics and control of epidemics stems from models that make highly simplifying assumptions or simulate contact networks from incomplete or proxy data. In another domain, the development of cybersecurity systems requires testing across diverse threat scenarios and validation across diverse network structures that are not yet known. In both examples, the systems of interest cannot be represented by a single exemplar network, but must instead be modeled as collections of networks in which the variation among them may be just as important as their common features. Such cases point to the importance of data-driven methods for synthesizing networks that capture both the essential features of a system and realistic variability in order to use them in such tasks as simulations and verification.

Because existing methods only reproduce a limited set of specified network properties, we introduce a novel strategy for synthesizing artificial networks, namely,

the multiscale entropic network generation (MUSKETEER, http://www.cs.clemson.edu/~isafro/musketeer)
We create a hierarchy of coarse networks; but, in contrast to the multiscale methods for computational and optimization problems, we do not optimize anything but *edit* the network at all scales of coarseness. During the editing process we allow only local changes in the network. The editing problem (or generation) is formulated and solved at all scales where primitives at the coarse scale (such as coarse nodes and edges) represent aggregates of primitives and their fractions at previous finer scale. Analogous to multiscale methods for computational problems, by using appropriate coarsening we are able to detect and use the geometry behind the original network at multiple scales, which can be interpreted as an additional property that is not captured by other network generation methods. It is known that the topology of many complex networks is hierarchical and thus might be produced through iterations of generative laws at multiple scales. In general, such generative laws often can be different at different scales, as evidenced by the finding that complex networks are self-dissimilar across scales. For example, the number of triangles in the original network that one can be interested in can be completely different from the number of triangles at coarse scales. These differences can naturally be reflected in the proposed multiscale framework.

To evaluate the performance of MUSKETEER, we generated a large number of networks and compared them to the real-world networks (from epidemiology, social science, etc.) for a variety of local and global structural properties (such as clustering, modularity, and degree distribution). For most properties, the generated ensemble yields a median value close to the original value and range of values that is fairly symmetric about the median.