JMLR: Workshop and Conference Proceedings vol 23 (2012) 45.1–45.3 25th Annual Conference on Learning Theory

Open Problem: Learning Dynamic Network Models from a Static Snapshot

Jan Ramon JAN.RAMON@CS.KULEUVEN.BE
Constantin Comendant CONSTANTIN.COMENDANT@CS.KULEUVEN.BE
Computer Science Dept., KU Leuven

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

In this paper we consider the problem of learning a graph generating process given the evolving graph at a single point in time. Given a graph of sufficient size, can we learn the (repeatable) process that generated it? We formalize the generic problem and then consider two simple instances which are variations on the well-know graph generation models by Erdős-Rényi and Albert-Barabasi.

1. Problem

An (undirected) graph is a pair $G = (V(G), E(G))$ with $V(G)$ a set of vertices and $E(G) \subseteq \big\{\{u, v\} \mid u, v \in V(G)\big\}$ a set of edges. The neighbors of a vertex v is the set $N_G(v) = \{u \in V(G) \mid \{u, v\} \in E(G)\}$. A (vertex-)labeled graph is a triple $G = (V(G), E(G), \lambda_G)$, where $(V(G), E(G))$ is a graph and $\lambda_G : V(G) \to \Sigma$ is a labeling function assigning to every vertex of G a label from an alphabet Σ. In the sequel of this paper, all graphs will be vertex-labeled.

A preferential attachment function is a function f mapping a vertex label $l \in \Sigma$, a labeled graph G and a subset $A \subseteq V(G)$ of its vertices to a positive real number. Here we will consider factored preferential attachment functions f for which there are functions f_l, f_d and f_v such that $f(l, G, A) = f_l(l) f_d(|A|) \prod_{v\in a} f_v(l, G, v) / \sum_A f_d(|A|) \prod_{v\in a} f_v(l, G, v)$.

A graph generation process according to the preferential attachment function f is a random sequence of graphs $\{G^{(t)}\}_{t=0}^{\infty}$ where $G^{(0)}$ is some starting graph, $V(G^{(t)}) = V(G^{(t-1)}) \cup \{v_t\}$ and $E(G^{(t)}) = E(G^{(t-1)}) \cup N_{G^{(t)}}(v_t)$ for $t \geq 1$, and $G^{(t)}$ is drawn with probability
$P(G^{(t)} \mid G^{(t-1)}) = f(\lambda_{G^{(t)}}(v), G^{(t-1)}, N_{G^{(t)}}(v_t)) / \sum_{l\in\Sigma} \sum_{A\subseteq V(G^{(t-1)})} f(l, G^{(t-1)}, A)$.

In this framework, one can generate Albert-Barabási (Barabási and Albert, 1999) random graphs, where new vertices attach to existing vertices with a probability proportional to the degree of these existing vertices, by setting $|\Sigma| = 1$, $f_l^{ab}(l) = 1$, $f_d^{ab}(x) = \delta(m, x)$ and $f_v^{ab}(l, G, v) = deg(v)$. Here, m is a constant and $\delta(m, x) = 1$ if $m = x$ and $\delta(m, x) = 0$ otherwise. One can generate Erdős-Rényi (Bollobás, 2001) random graphs setting $|\Sigma| = 1$, $f_l^{er}(l) = 1$, $f_d^{er}(x) = 1$ and $f_v^{er}(l, G, v) = 1$.

In the context of our problem, a hypothesis space is a set of preferential attachment functions. The problem of learning dynamics from a snapshot is the problem where we are given a hypothesis space H, a loss function $L : H \times H \to \mathsf{R}^+$, an integer t, and a graph $G^{(t)}$ generated according to some unknown $f^* \in H$. The task is then to find a $\hat{f} \in H$ minimizing $L(f^*, \hat{f})$. A learning bound for a hypothesis space H and loss function L is an algorithm A and a polynomial $p(\cdot, \cdot)$ such that for any ϵ and δ, if $t \geq p(1/\epsilon, 1/\delta)$ and a graph $G^{(t)}$ is generated using $f^* \in H$ and given as input to algorithm A, there holds $P(L(f^*, A(G^{(t)})) < \epsilon) > 1 - \delta$, i.e. with high confidence $\hat{f}$ closely approximates f^*. Note that the probability is over the random graph generation process, and independent of the prior distribution of f^*.

2. Related work and motivation

There are several distinct lines of research associated with large networks and random graphs, which we will mention briefly.

Probably the most common and well-established area of statistical study of networks is the study of their asymptotic properties given certain parameters of their formation. The classic example in this respect is the seminal work by Erdős and Rényi on the random graphs (Erdős and Rényi, 1959), but also more recent models for small-world (Watts and Strogatz, 1998) or scale-free (Barabási and Albert, 1999) networks.

A different and less theoretical approach is to simulate the evolution of the network in order to be able to make predictions about trends (Edmonds et al., 2007).

Yet another area of research is the problem of learning the dynamics of a network given temporal data (also called longitudinal data). However, in many real-world settings, temporal data is not available or at least much harder to obtain than a snapshot of the current state of a network. E.g. even though nobody logged all events in the construction of the Internet, one can easily access the current state of this network.

To the best of our knowledge, learning the dynamics of a network given a snapshot at a particular point in time has not been studied yet thoroughly. Therefore, in this paper, we want to raise the question for which classes of complex generative models one can provide learning guarantees. In the next section, we provide two simple cases, one of which can be solved easily.

3. Examples

Linking to similar vertices Consider graphs whose vertices are assigned a boolean label, i.e. $\lambda : G \to \Sigma = \{0, 1\}$. In this model, intuitively, new vertices attach preferentially to vertices with the same label. Therefore, let $0 \leq w \leq 1$ express the importance of attaching to similarly labeled vertices, and $0 \leq p \leq 1$ express the probability that a new vertex gets label 1. We define the preferential attachment function as follows. First, set $f_l^b(1) = p$ and $f_l^b(0) = 1 - p$. Let each new vertex attach to exactly m existing vertices, i.e. $f_d(d) = \delta(m, d)$. Finally, let $f_v^b(l, G, v) = 1 + w.\delta(l, \lambda(v))$ so that new vertices have a preference to attach to vertices with the same label. Let us consider as the loss function the difference between the real and the learned parameters, i.e. $L((p^*, w^*), (\hat{p}, \hat{w})) = |p^* - \hat{p}| + |w^* - \hat{w}|$. We can now estimate p and w. In the following derivation, for simplicity of explanation we will make abstraction of the small fraction of vertices and edges that were already present in $G^{(0)}$, it is easy to take these into account. Consider the graph $G = G^{(t)}$. First, the number of vertices with label 1 in G follows a binomial distribution. We can estimate $\hat{p} = |\{v \in V(G)|\lambda_G(v) = 1\}|/|V(G)|$ and $var(p - \hat{p}) = p(1-p)/|V(G)|$ gives us the confidence of our estimate. Moreover, we can count the number $E_{x,y}$ of edges whose endpoints are labeled x and y. Let $E = E_{0,0} + E_{0,1} + E_{1,1}$. The fraction of new edges with endpoints 0 and 0 is approximately the probability $1 - p$ of getting a new vertex with label 0, multiplied with the probability $(1-p)(1+w)/(1+w-pw)$ of attaching it to a vertex of label 0, i.e. $E_{0,0}/E = (1-p)^2(w+1)/(1+w-pw)$. Similarly $E_{1,1}/E = p^2(w+1)/(pw+1)$ and

$E_{0,1} = p(1-p)/(pw+1) + (1-p)p/(1+w-pw)$. From these several equations one can solve w and put confidence bounds on it.

A labeled Barabási-Albert model For another example, consider a different network generation process: the vertices of the graph can be assigned two labels, 0 and 1 with probability $(1-p)$ and p as above (i.e. $f_l^{lb}(1) = p$ and $f_l^{lb}(0) = 1-p$. Let $w > 0$. Let $f_d^{lb}(d) = \delta(m,d)$ ensure that every new vertex is attached to exactly m existing vertices. Let $N_G^l(v) = \{u \in N_G(v) \mid \lambda(u) = l\}$. Then, $f_v^{lb}(l,G,v) = N_G^0(v) + w.N_G^1(v)$. Just as in the Barabasi-Albert model, the probability of attachment to existing vertices is proportional to the degree of these existing vertices, but here a 'weighted' degree is used where vertices with label 1 have a weight w instead of a weight 1. Contrarily to the previously discussed case, one can show that asymptotically $(E_{0,0}/E, E_{1,1}/E, E_{0,1}/E) = ((1-p)^2, p^2, 2\,p)$ is not a function of w, i.e. the weight w that is part of the graph generation process cannot be computed directly from the edge-endpoint-pair distribution, nor the vertex label distribution of the graph.

4. Conclusion

Solving the problem of learning dynamics (a preferential attachment function) from a snapshot of the evolving network at a particular point in time, will probably involve methods used to study the asymptotic properties of such networks, augmented with an analysis of the confidence of the estimations of the model parameters. To the best of our knowledge, the amount of work in this direction is limited.

Acknowledgements

This work was supported by ERC Starting Grant 240186 'MiGraNT: Mining Graphs and Networks, a Theory-based approach'.

References

A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.

B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.

B. Edmonds, C. Hernandez, and K. G. Troitzsch. *Social Simulation: Technologies, Advances and New Discoveries*. IGI Publishing, 2007.

P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684): 440–442, 1998.