2018-07-03

1 Graphs and linear algebra

Formally, an unweighted graph is $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Informally, $\mathcal{V}$ consists of things we want to model and $\mathcal{E}$ represents the relations between them. It is a very flexible representation: we use graphs to represent friendships between people, wires between routers, citations between papers, links between objects in a data structure, and many other things. When the bare topology of the relationships does not provide enough modeling power, we might also consider including functions on $\mathcal{V}$ or $\mathcal{E}$ corresponding to different attributes. The most common case is a scalar *weight* function assigned to each edge that corresponds to the importance of the relation: in a social network, for example, maybe a close and active friendship has more weight than a casual acquaintance. We use relationship information encoded in graphs to reason about logical groupings (whether we call them communities or clusters), about power relations and influence, and about dynamic processes like the spread of rumors or disease.

Matrix methods for network analysis rely on a sort of pun: we encode the network as a matrix, translate the question into linear algebraic terms[1], and commence to compute. There are many possible matrices associated with a graph, and we use them to reason about different things. We consider three interpretations of these network matrices:

- A matrix may represent a *linear map* between *different spaces*, typically mapping vertex properties to edge properties, or vice-versa. Examples include the discrete gradient operator and the edge sum operator.

- A matrix may represent an *operator* mapping the space of functions over the vertices (or over edges) to itself. Examples include transition matrices for random walks defined on the graph.

- A matrix may represent a *quadratic form* mapping functions on the vertices (or edges) into scalars. Often the quadratic form has an easy to interpret meaning for special inputs; for example, the quadratic form for the *combinatorial Laplacian* counts cut edges in graph partitioning.

[1] "Mathematicians are like Frenchmen: whatever you say to them they translate into their own language and forthwith it is something entirely different." – Goethe

2 Adjacency and degrees

If we identify vertices in the graph with indices $1, \ldots, n$, then the (unweighted) *adjacency matrix* $A \in \mathbb{R}^{n \times n}$ has entries

$$a_{ij} = \begin{cases} 1, & (i,j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases}$$

For graphs in which edges have positive weights, we sometimes use the *weighted* adjacency matrix, with a_{ij} giving the edge weight for $(i,j) \in \mathcal{E}$.

The *degree* d_i of a node is the sum of the weights on the incident edges. When the graph is directed, the in-degree and out-degree may differ; for the moment, we will stick with the directed case. We let $d \in \mathbb{R}^n$ denote the vector of weighted node degrees, and let D denote the matrix in which the weighted node degrees appear on the diagonal.

The adjacency matrix and the degree matrices are building blocks for several other matrices, but they are also useful on their own. First, as a linear operator, the adjacency matrix accumulates values from neighbors; that is,

$$(Ax)_i = \sum_{j \in N_i} x_j$$

where $N_i = \{j : (i,j) \in \mathcal{E}\}$ is the neighborhood of i. If x_j is the number of paths of length k leading from starting point to node j, then $(Ax)_i$ is the number of paths of length k to all neighbors of node i — that is, the total number of paths of length $k+1$ to node j. Therefore,

$$[A^k]_{ij} = \text{number of paths of length } k \text{ from } i \text{ to } j.$$

We use this formula in many ways; for example, it lets us write number of triangles in an undirected graph (closed cycles of length three) as

$$\text{number of triangles} = \frac{1}{3} \sum_i [A^3]_{ii} = \frac{1}{3} \operatorname{tr}(A^3)$$

where we divide by three because each triangle is counted once for each of its vertices. We also know, for example, how to approximate the number of long paths between i and j in terms of the dominant eigenvector (and associated eigenvalue) of A.

The matrix A also defines a quadratic form that is useful for counting edges. Let $x \in \{0,1\}^n$ be the indicator for a subset of vertices $S \subset \mathcal{V}$. Then

$$x^T A x = \sum_{i,j} a_{ij} x_i x_j = \sum_{(i,j) \in S \times S} a_{ij},$$

i.e. $x^T A x$ is the total (directed) edge weight between nodes in S, or twice the total undirected edge weight. If x is an indicator then $x^T x = |S|$, and so

$$\frac{x^T A x}{x^T x} = \text{mean degree within } S.$$

If S is a clique, the mean degree within S is $|S| - 1$; therefore

$$|S| - 1 = \frac{x^T A x}{x^T x} = \rho_A(x) \leq \lambda_{\max}(A),$$

since $\lambda_{\max}(A)$ is the largest possible value for the Rayleigh quotient $\rho_A(x)$. Hence, the maximum clique size $k(\mathcal{G})$ has the bound

$$k(\mathcal{G}) \leq 1 + \lambda_{\max}(A).$$

This is an example of a result in *spectral graph theory*, i.e. the study of graphs in terms of eigenvalues and eigenvectors of associated matrices. In fact, another continuous optimization problem due to Motzkin and Straus gives the clique number exactly:

$$1 - 1/k(\mathcal{G}) = \max_{x \in \Delta_n} x^T A x, \text{ where } \Delta_n \equiv \{x \in \mathbb{R}^n : x \geq 0, e^T x = 1\}.$$

The optimization is now carried out over the simplex Δ_n rather than over the Euclidean unit ball used in the spectral bound.

If x is an indicator for a set S, we can also use the quadratic form

$$x^T D x = \sum_{i \in S} d_i = \text{edges incident on } S.$$

Therefore,

$$x^T D x - x^T A x = x^T (D - A) x = \text{edges between } S \text{ and } S^C.$$

We will see more of the *combinatorial Laplacian* matrix $L = D - A$ shortly.

3 Random walks and normalized adjacency

Now consider a random walk on a $\mathcal{G}$, i.e. a Markov process where the walker location X^{t+1} at time $t+1$ is chosen randomly from among the neighbors of the previous location X^t with probability determined by the edge weights. Using the properties of conditional probability,

$$P\{X^{t+1} = i\} = \sum_j P\{X^{t+1} = i | X^t = j\} P\{X^t = j\},$$

and the rule for randomly choosing a neighbor gives

$$P\{X^{t+1} = i | X^t = j\} = \frac{a_{ij}}{d_j}.$$

Letting $\pi^{t+1} \in \mathbb{R}^n$ be the column vector whose entries represent the probability that $X^{t+1} = i$, and similarly with π^t, we write the equation for conditional probability concisely as

$$\pi^{t+1} = (AD^{-1})\pi^t.$$

The matrix $T = AD^{-1}$ is the *transition matrix* for the random walk Markov chain[2]. Powers of T have an interpretation similar to that of powers of A, but rather than counting length k paths, T^k computes probabilities:

$$[T^k]_{ij} = P\{X^k = i | X^0 = j\}.$$

Assuming the graph is connected and and aperiodic[3], the matrix T has a unique eigenvalue at 1, and all other eigenvalues are inside the unit circle. In this case,

$$\lim_{k\to\infty} T^k = T^\infty = (\pi^*)e^T$$

where π^* is a probability vector representing the *stationary distribution* for the Markov chain. In the undirected case, the stationary distribution is rather simple: $\pi^*_i = d_i/(2m)$. Things are more interesting for directed graphs.

[2]We use the convention that probability densities are column vectors, and that a_{ij} represents a transition from j to i. If $\mathcal{G}$ is directed, we also denote by D the out-degree of the nodes. This is consistent with the conventions in numerical linear algebra; in other areas, probability densities are typically rows.

[3]The graph is aperiodic if there is some k such that there is a length k path between any nodes i and j.

While the eigenvalue at 1 and the associated stationary distribution are particularly interesting, the other eigenvalues and vectors are also interesting. In particular, suppose $T = V\Lambda V^{-1}$, and consider

$$\|T^k - T^\infty\| = \|V\bar{\Lambda}^k V^{-1}\| \leq \kappa(V)|\lambda_2|^k,$$

where $\kappa(V) = \|V\|\|V^{-1}\|$ is the condition number of the eigenvector matrix, $\bar{\Lambda}$ is the diagonal matrix of eigenvalues with the eigenvalue at one replaced by zero, and $|\lambda_2|$ is the maximum modulus of all eigenvalues other than the eigenvalue at one. Therefore, the asymptotic rate of convergence of the Markov chain, also known as the *mixing rate* is determined by the second-largest eigenvalue modulus of T.

To understand the mixing rate in more detail (in the undirected case), it is helpful to consider the closely related *normalized adjacency matrix*

$$\bar{A} = D^{-1/2}AD^{-1/2}.$$

Note that $\bar{A} = D^{-1/2}TD^{1/2}$, so

$$(v, \lambda) \text{ an eigenpair of } \bar{A} \iff (D^{1/2}v, \lambda) \text{ an eigenpair of } T.$$

The eigenvalues of $\bar{A}$ are critical points of

$$\rho_{\bar{A}}(x) = \frac{x^T D^{-1/2} A D^{-1/2} x}{x^T x};$$

substituting $x = D^{1/2}y$, we have

$$\rho_{\bar{A}}(D^{1/2}y) = \rho_{(A,D)}(y) = \frac{y^T A y}{y^T D y}.$$

If y is an indicator for a set S, this last expression represents the fraction of edges incident on S that are to other nodes in S. If $z = 2y - e$ is $+1$ on S and -1 on S^c, then $z^T D z = 2m$ and $z^T A z$ is $2m$ minus twice the total weight $|C(S)|$ of edges from S to S^C; hence,

$$\frac{z^T A z}{z^T D z} = 1 - \frac{|C(S)|}{m}$$

If we restrict to the case where the same number of edges are incident on S and S^C, then z is D-orthogonal to the all one vector, and so $\rho_{(A,D)}(z)$ is a lower bound on the eigenvalue closest to one. Thus, spectral analysis lets us bound the mixing rate in terms of the normalized cut size $|C(S)|/m$.

4 Discrete gradients and the Laplacian

For an unweighted graph, the *discrete gradient* $G \in \mathbb{R}^{m \times n}$ is a matrix in which each row represents an edge $(i,j) \in \mathcal{E}$ by $(e_i - e_j)^T$. If x is an indicator for a set S, then Gx is nonzero (± 1) only on edges between S and S^c; hence,

$$\|Gx\|^2 = \text{edges between } S \text{ and } S^C.$$

We can rewrite this as $x^T G^T G x = x^T L x$ where

$$L = \sum_{(i,j)\in\mathcal{E}} (e_i - e_j)(e_i - e_j)^T.$$

Each term in this sum contributes one to the l_{ii} and l_{jj} entries through the $e_i e_i^T$ and $e_j e_j^T$ products; the cross terms fill in $l_{ij} = l_{ji} = -1$. Putting everything together, we have

$$L = D - A.$$

The matrix L is known as the *combinatorial Laplacian*, or sometimes simply as the Laplacian. The same construction holds in the weighted case, where it corresponds to

$$L = \sum_{(i,j)\in\mathcal{E}} a_{ij}(e_i - e_j)(e_i - e_j)^T$$

where a_{ij} is the weight of the (i,j) edge. In either case, the smallest eigenvalue of L is zero, corresponding to an eigenvector of all ones. The multiplicity of the zero eigenvalue is equal to the number of connected components in the graph; assuming there is only one connected component, the second largest eigenvalue λ_2 is a lower bound on $x^T L x$ for any $x^T e = 0$; if we choose x to be a ± 1 vector indicating the split between equal size sets S and S^c, then $x^T L x$ also gives four times the number of edges cut by the partitioning. Hence, $\lambda_2/4$ is a lower bound on the minimal bisector size.

The combinatorial Laplacian also can be interpreted as a linear operator, and in this guise it plays a role as the generator for a *continuous-time random walk* involving a random walk in which the time elapsed between each consecutive pair of steps is given by an independent exponential random variable with mean one. In this case, we have

$$\exp(-sL)_{ij} = P\{X(s) = i | X(0) = j\}.$$

The matrix $\exp(-sL)$ is known as the *heat kernel* on the graph because it can also describe continuous-time diffusion of heat on a graph.

The *normalized* Laplacian is $\bar{L} = D^{-1/2}LD^{-1/2} = I - \bar{A}$; the eigenvalues of the normalized Laplacian can also be expressed as critical points of the generalized Rayleigh quotient

$$\rho_{(L,D)}(x) = \frac{x^T L x}{x^T D x}.$$

Thus twice the Rayleigh quotient gives the fraction of all edges that go between S and S^C if x is a vector which is $+1$ on the set S and -1 on S^c.

5 Discrete sums and the signless Laplacian

The *discrete sum* operator is $G^+ \in \mathbb{R}^{m\times n}$ where each row corresponds to an edge $(i,j) \in \mathcal{E}$ and has the form $(e_i + e_j)^T$. The *signless Laplacian* is

$$L^+ = D + A = (G^+)^T(G^+) = \sum_{(i,j)\in\mathcal{E}} (e_i + e_j)(e_i + e_j)^T.$$

The signless Laplacian is positive semi-definite; but unlike the combinatorial Laplacian, it may or may not have a null vector. If there is a null vector x for the signless Laplacian, then we must have $x_i = -x_j$ whenever $(i,j) \in \mathcal{E}$; this implies that x indicates *bipartite structure* where the set S with positive elements can have edges to a set S' with negative elements, but neither S nor S' may have any other edges. There has been some work on the spectral theory for the signless Laplacian, but it is generally much less used than the combinatorial Laplacian.

6 Modularity matrix

Suppose we want to find a tight cluster in our graph. One approach is to look for sets of nodes that make $x^T A x$ large; but large relative to what? For an undirected graph, we use the quadratic form $x^T A x$ to count internal edges in a set. But a set may have many internal edges purely as an accident of having many high-degree nodes: high-degree nodes are highly likely to connect to each other simply because they are highly likely to connect to

anyone! Hence, we would like a reference model so that we can see whether the edge density in a subgraph is "unusually" high relative to that reference.

The simplest reference model that takes into account the effect of degree distribution is sometimes called the *configuration model*. In the configuration model, we prescribe the vector d of expected node degrees. We then add m edges, each of which is (i, j) with probability $d_i d_j/(2m)^2$; self-loops and repeating edges are allowed. With this construction, the expected adjacency is

$$\bar{A} = \frac{dd^T}{2m},$$

which has the correct expected degree distribution. The *modularity* matrix is defined to be

$$B = A - \bar{A},$$

and if x is an indicator for S, the quadratic form $x^T Bx$ indicates the number of "excess edges" within S compared to what we would predict in the configuration model.

7 And many more

The list of graph matrices that we have discussed is by no means exhaustive. We will see a few more examples this week, including the heat kernel matrix and the PageRank matrix. But there are many more besides; for example, recent work by Leskovec and Benson used a *motif adjacency matrix* $M = (A \odot A)A$ for which m_{ij} represents the number of triangles in the graph involving the edge (i, j). And one can define ever more exotic matrix representations. However, the adjacency, Laplacian, and their close neighbors suffice for many applications.