Modularity of Erdős-Rényi random graphs

Colin McDiarmid

Oxford University

Martin Dyer Day, QMUL, 16 July 2018

joint work with Fiona Skerman

Modularity and communities

Modularity was introduced by Newman and Girvan in 2004 to give a measure of how well a graph can be divided into communities.

It now forms the backbone of the most popular algorithms used to cluster real data, with many applications, from protein discovery to identifying connections between websites.

See for example surveys by Fortunato (2010), and Porter Onnela and Mucha (2009), on the use of modularity for community detection in networks.

Definition of modularity

Let $G = (V, E)$ be a graph with $m \geq 1$ edges. For a set A of vertices, let $e(A)$ be the number of edges within A, and let $\mathrm{vol}(A)$ be the sum over the vertices $v \in A$ of the degree d_v.

Given a partition $\mathcal{A}$ of V, the modularity of $\mathcal{A}$ on G is

$$\begin{aligned} q_{\mathcal{A}}(G) &= \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \left(\mathbf{1}_{uv \in E} - \frac{d_u d_v}{2m} \right) \\ &= \frac{1}{m} \sum_{A \in \mathcal{A}} e(A) - \frac{1}{4m^2} \sum_{A \in \mathcal{A}} \mathrm{vol}(A)^2; \end{aligned}$$

and the modularity of G is $q^*(G) = \max_{\mathcal{A}}(G)$.

Isolated vertices are irrelevant; and we shall not consider empty graphs (that is, with no edges).

modularity: understanding the definition

$$q_{\mathcal{A}}(G) = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \left(\mathbf{1}_{uv \in E} - \frac{d_u d_v}{2m} \right).$$

If we pick uniformly at random a multigraph with the same degrees as G, then the expected number of edges between vertices u and v is essentially

$$\frac{d_u d_v}{2m}.$$

This is the rationale for the definition: whilst rewarding the partition for capturing edges within the parts, we should penalise by the expected number of edges.

edge-contribution and degree tax

The second equation

$$q_{\mathcal{A}}(G) = \frac{1}{m}\sum_{A\in\mathcal{A}} e(A) - \frac{1}{4m^2}\sum_{A\in\mathcal{A}} \mathrm{vol}(A)^2$$

expresses $q_{\mathcal{A}}(G)$ as the difference of two terms:

the edge contribution $q^E_{\mathcal{A}}(G) = \frac{1}{m}\sum_A e(A)$,

and the degree tax $q^D_{\mathcal{A}}(G) = \frac{1}{4m^2}\sum_A \mathrm{vol}(A)^2$.

Since $q^E_{\mathcal{A}}(G) \le 1$ and $q^D_{\mathcal{A}}(G) > 0$, we have $q_{\mathcal{A}}(G) < 1$. Also, the trivial partition $\mathcal{A}_0$ with one part has $q^E_{\mathcal{A}_0}(G) = q^D_{\mathcal{A}_0}(G) = 1$, so $q_{\mathcal{A}_0}(G) = 0$. Thus

$$0 \le q^*(G) < 1.$$

Modularity: some examples

(a) Let G be a tree with m edges and max degree $\Delta = o(m)$. Then $q^*(G) = 1 - o(1)$. (True also if *treewidth* $\cdot\, \Delta = o(m)$.)

(b) Let G be an m-edge subgraph of the square lattice. Then $q^*(G) = 1 - o(1)$.

(c) $q^*(K_n) = 0$ (and indeed $q^*(G) = 0$ if G is K_n less at most $n/2$ edges).

(d) If G consists of $k \geq 1$ cliques all of the same size, then

$$q^*(G) = 1 - 1/k.$$

Two theorems on the modularity of random graphs $G(n,p)$

Here are two theorems of ours (McD and Skerman) on $q^*(G(n,p))$. First, the overview.

Theorem (3 phases theorem)

(a) *If* $n^2 p \to \infty$ *and* $np \leq 1 + o(1)$ *then* $q^*(G(n,p)) \xrightarrow{p} 1$.

(b) *Given* $1 < c_0 \leq c_1$, *there exists* $\delta > 0$ *such that, if* $c_0 \leq np \leq c_1$, *then whp* $\delta < q^*(G(n,p)) < 1 - \delta$.

(c) *If* $np \to \infty$ *then* $q^*(G(n,p)) \xrightarrow{p} 0$.

To prove part (a) it suffices to consider the partition into components. Part (c) and much of part (b) follow from the next theorem.

Two theorems on $q^*(G(n,p))$

Theorem (the $(np)^{-1/2}$ theorem)

There exist $0 < a < b$ such that, if $np \geq 1$ and $p \leq 0.99$, then

$$\frac{a}{\sqrt{np}} < q^*(G(n,p)) < \frac{b}{\sqrt{np}} \quad \textit{whp.}$$

This confirms a conjecture in 2006 by Reichardt and Bornholdt (and refutes another conjecture from the physics literature).

The upper bound may be proved using the expander mixing lemma (not here).

The lower bound follows by considering a simple algorithm *Swap* (or, for a more limited range of p, from recent work on stochastic block models.)

Two theorems on $q^*(G(n,p))$

As we noted, much of the $np > 1$ part of the 3 phases theorem follows from the $(np)^{-1/2}$ theorem.

To complete the proof for $np > 1$, we need to show that $q^*(G(n,p)) < 1 - \delta$ whp when np is just above 1.

To do this, we may use the result that whp, splitting the giant component roughly into halves must break $\Omega(n)$ edges (Luczak and McD 2001).

Swap gives the $(np)^{-1/2}$ lower bound

Given a graph G, the algorithm *Swap* runs in linear time and yields a balanced bipartition $\mathcal{A}$ of the vertices.

Theorem

There are constants c_0 and $a > 0$ such that if $p = p(n)$ satisfies $1 \leq np \leq n - c_0$, then whp

$$q_{\mathcal{A}}(G_{n,p}) \geq a\left(\frac{1-p}{np}\right)^{1/2};$$

and if also $np \geq c_0$ we may take $a = \frac{1}{5}$.

Idea of *Swap*

The algorithm *Swap* starts with a balanced bipartition of the vertex set into $A \cup B$, which has modularity very near 0 whp.

By swapping some pairs between A and B, whp we can increase the edge contribution significantly, without changing the distribution of the degree tax (and without introducing dependencies which would be hard to analyse).

The algorithm *Swap*

Assume for simplicity that $6|n$ and write $n = 6k$. Start with the bipartition $\mathcal{A}$ of $V = [n]$ into $A = \{j \in V : j \text{ is odd}\}$ and $B = \{j \in V : j \text{ is even}\}$. Whp $q_{\mathcal{A}}(G_{n,p})$ is very close to 0.

Let $V_0 = [4k]$, let $V_1 = \{4k+1, \ldots, 6k\}$. Let $A_0 = A \cap V_0$, $A_1 = A \cap V_1$ and $B_0 = B \cap V_0$, $B_1 = B \cap V_1$. The four sets A_i, B_i are pairwise disjoint with union V.

Initially V_0 is partitioned into $A_0 \cup B_0$: the algorithm *Swap* 'improves' this partition, keeping A_1, B_1 fixed. For $i = 1, \ldots, 2k$ let $a_i = 2i - 1$ and $b_i = 2i$, so $A_0 = \{a_1, \ldots, a_{2k}\}$ and $B_0 = \{b_1, \ldots, b_{2k}\}$. We improve the partition $V_0 = A_0 \cup B_0$ is by independently swapping a_i and b_i for certain values i.

T_i and swapping a_i, b_i

For each $i \in [2k]$ let

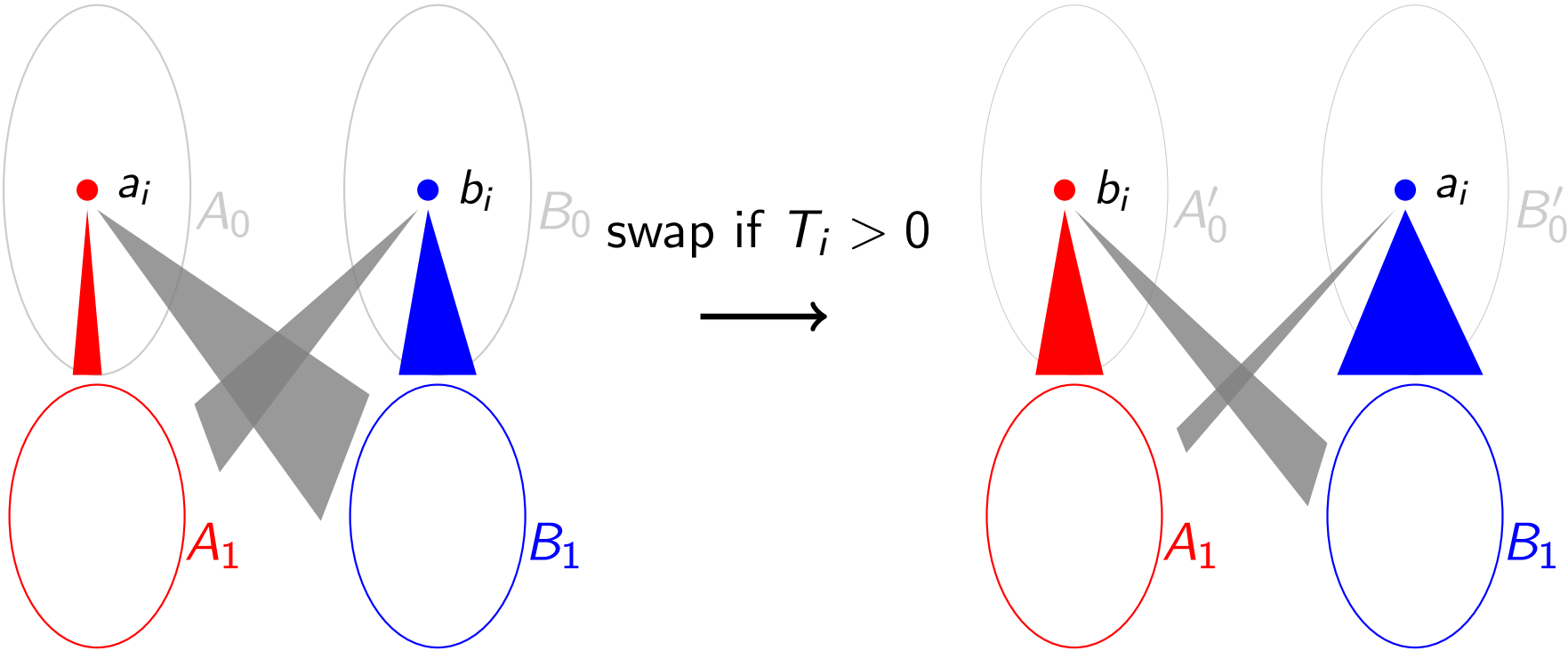
$$T_i = e(a_i, B_1) - e(a_i, A_1) + e(b_i, A_1) - e(b_i, B_1).$$

The random variables $T_1, \ldots, T_{2k}$ are iid.

Also $\mathbb{E}[T_i] = 0$, $\mathrm{var(T_i)} = 4\mathrm{kp}(1-\mathrm{p})$; and $\mathbb{E}[|T_i|] = \Theta((np(1-p))^{1/2})$.

If $T_i > 0$ and we swap a_i and b_i between A_0 and B_0 (that is, replace A_0 by $(A_0 \setminus \{a_i\}) \cup \{b_i\}$ and similarly for B_0) then $e(A, B)$ decreases by T_i, so the edge contribution of the partition increases.

Illustration of swapping



T^* and swaps

Swap makes all such swaps (looking only at possible edges between V_0 and V_1), yielding the balanced bipartition $\mathcal{A}' = (A', B')$, where $A' = A'_0 \cup A_1$ and $B' = B'_0 \cup B_1$.

Let $T^* = \sum_{i \in [2k]} |T_i|$. Then

$$e(A'_0, A_1) + e(B'_0, B_1) - (e(A'_0, B_1) + e(A_1, B'_0)) = T^*.$$

But $e(A'_0, A_1) + e(B'_0, B_1) + (e(A'_0, B_1) + e(A_1, B'_0)) = e(V_0, V_1)$, so

$$e(A'_0, A_1) + e(B'_0, B_1) = \tfrac{1}{2} e(V_0, V_1) + \tfrac{1}{2} T^*.$$

T^* and edge contribution

T^* is the sum of the $2k \approx n/3$ iid random variables $|T_i|$, so whp

$$T^* \approx 2k\,\mathbb{E}[|T_1|] = \Theta(n^{3/2}(p(1-p))^{1/2}).$$

Thus whp the edge contribution for $\mathcal{A}'$ beats that for the initial bipartition $\mathcal{A}$ by

$$\Theta\left(\frac{n^{3/2}(p(1-p))^{1/2}}{n^2 p}\right) = \Theta\left(\left(\frac{1-p}{np}\right)^{1/2}\right).$$

In other words

$$q^E_{\mathcal{A}'}(G_{n,p}) - q^E_{\mathcal{A}}(G_{n,p}) = \Theta\left(\left(\frac{1-p}{np}\right)^{1/2}\right) \quad \text{whp.}$$

What about degree tax?

Our decisions about when to swap are symmetric. In the two cases

$$e(a_i, B_1) = w, e(a_i, A_1) = x \text{ and } e(b_i, A_1) = y, e(b_i, B_1) = z$$

$$e(a_i, B_1) = y, e(a_i, A_1) = z \text{ and } e(b_i, A_1) = w, e(b_i, B_1) = x.$$

we make the same decision (swap iff $w-x + y-z > 0$). It follows that the degree tax for $\mathcal{A}'$ has exactly the same distribution as for $\mathcal{A}$. We find

$$q^D_{\mathcal{A}'}(G_{n,p}) - q^D_{\mathcal{A}}(G_{n,p}) = o\left((\frac{1-p}{np})^{1/2}\right) \text{ whp.}$$

completing the *Swap* story

Putting together the results on edge contribution and on degree tax we find

$$q_{\mathcal{A}'}(G_{n,p}) - q_{\mathcal{A}}(G_{n,p}) = \Theta\left((\frac{1-p}{np})^{1/2} \right) \quad \text{whp.}$$

But whp $q_{\mathcal{A}}(G_{n,p})$ is very near 0, and so

$$q_{\mathcal{A}'}(G_{n,p}) = \Theta\left((\frac{1-p}{np})^{1/2} \right) \quad \text{whp}$$

as required.

around $np = 1$

We saw that $q^*(G_{n,p}) = 1 + o(1)$ whp if $np \leq 1 + o(1)$ (assuming that $n^2 p \to \infty$);

and $q^*(G_{n,(1+\epsilon)/n})$ is bounded below 1 for fixed $\epsilon > 0$.

Recently we have found: $q^*(G_{n,(1+\epsilon)/n}) = 1 - \Theta(\epsilon^2)$ whp.

thanks for your attention

and

Happy Birthday Martin!