Geometry of Exponential Random Graph Models

Ha Khanh Nguyen
The Ohio State University

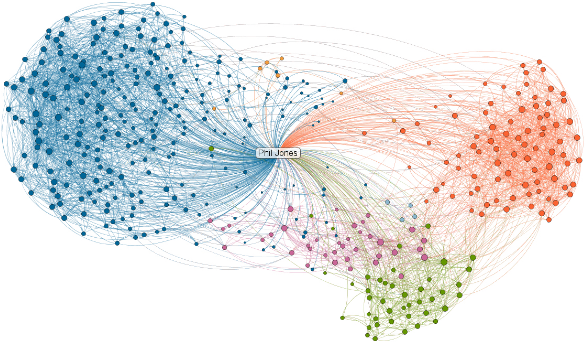
Elizabeth Gross
University of Hawaii, Manoa

Christopher O'Neill
University of California, Davis

Rodolfo Garcia
San José State University

Big Statistical Challenge: *Researchers want to know how well a proposed model fits an observed network.*

- One of the most studied classes of network models are exponential random graph models, which are defined by network statistics (sufficient statistics).



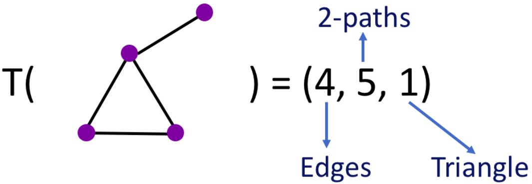
Exponential Random Graph Models

The model:

$$P_\theta(G) = Z(\theta)e^{\theta \cdot T(G)}, \ \theta \in \mathbb{R}^d$$

- $\theta \in \mathbb{R}^d$ is the **parameter vector**.
- $Z(\theta)$ is the **normalizing constant**.
- $T : \{\text{Graphs with } n \text{ vertices}\} \to \mathbb{Z}^d_{\geq 0}$ is the **sufficient statistic**.

Example to have in mind: Let $T(G)$ count the number of edges, 2-paths and triangles of the graph G.



Background and Notation

Adjacency Matrix: *Let $G = (V, E)$ be a graph with n vertices. Let A be the adjacency matrix of the graph, i.e.*

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

Network Statistics: *The sufficient statistics of the edge-2path-triangle model are:*

- The number of edges of G: $t_1 = \sum_{i<j} A_{ij}$.
- The number of 2-paths in G: $t_2 = \sum_{i<j} (A^2)_{ij}$.
- The number of triangles in G: $t_3 = \frac{1}{6}\mathrm{tr}A^3$.

Adjacency Matrix: *Let $G = (V, E)$ be a graph with n vertices. Let A be the adjacency matrix of the graph, i.e.*

$$A_{ij} = \begin{cases} x_{ij} & \text{if } i < j, \\ 0 & \text{if } i = j, \\ x_{ji} & \text{if } i > j. \end{cases} \quad (2)$$

Network Statistics: *The sufficient statistics of the edge-2path-triangle model are:*

- The number of edges of G: $t_1 = \sum_{i<j} A_{ij}$.
- The number of 2-paths in G: $t_2 = \sum_{i<j} (A^2)_{ij}$.
- The number of triangles in G: $t_3 = \frac{1}{6}\text{tr}A^3$.

Mantra of Algebraic Statistics

Statistical Models ⟷ Algebraic Varieties

Statistics | **Algebraic Geometry**

Algebraic Ideal

Definition

$k[x_1, \ldots, x_n]$ is the set of all polynomials in $x_1, \ldots, x_n$ with coefficients in k where k is a field.

Note: The field k can be $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$.
Example: $x + y - z^2$ is a polynomial in $\mathbb{R}[x, y, z]$.

Definition

A subset $I \subset k[x_1, \ldots, x_n]$ is an **ideal** if it satisfies:

1. $0 \in I$.
2. If $f, g \in I$, then $f + g \in I$.
3. If $f \in I$ and $h \in k[x_1, \ldots, x_n]$, then $hf \in I$.

Algebraic Ideal

Definition

Let $f_1, \ldots, f_s$ be polynomials in $k[x_1, \ldots, x_n]$. Then we set

$$\langle f_1, \ldots, f_s \rangle = \left\{ \sum_{i=1}^{s} h_i f_i : h_1, \ldots, h_s \in k[x_1, \ldots, x_n] \right\}.$$

Lemma

If $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, then $\langle f_1, \ldots, f_s \rangle$ is an ideal.
We will call $\langle f_1, \ldots, f_s \rangle$ the **ideal generated by** $f_1, \ldots, f_s$.

Example: $I = \langle x + y, x^2 - y + z^3 \rangle$ is an ideal of $\mathbb{R}[x, y, z]$ and it is generated by $x + y$ and $x^2 - y + z^3$.

Algebraic Variety

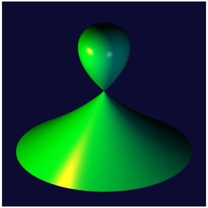
Definition

Let $f_1, \ldots, f_s$ be polynomials in $k[x_1, \ldots, x_n]$. Then we set

$$\mathbf{V}(f_1, \ldots, f_s) = \{(a_1, \ldots, a_n) \in k^n : f_i(a_1, \ldots, a_n) = 0 \ \forall 1 \le i \le s\}.$$

We call $\mathbf{V}(f_1, \ldots, f_s)$ the **affine variety** defined by $f_1, \ldots, f_s$.

Example:



$\mathbf{V}(x^2 + y^2 + z^3 - z^2)$

$\mathbf{V}((y^2 + z^2 - 1)^2 + (x^2 + y^2 - 1)^3)$

https://homepage.univie.ac.at/herwig.hauser/bildergalerie/gallery.html

Algebraic Variety

Interpretation: An affine variety $\mathbf{V}(f_1, \ldots, f_s) \subset k^n$ is the set of all solutions of the system of equations

$$f_1(x_1, \ldots, x_n) = \cdots = f_s(x_1, \ldots, x_n) = 0.$$

Similarly, given a polynomial ideal $\mathbf{I} = \langle g_1, \ldots, g_s \rangle \subset k[x_1, \ldots, x_n]$, $\mathbf{V}(\mathbf{I})$ is an affine variety in k^n. It is the set of all solutions of the system of equations where all the polynomials in $\mathbf{I}$ set equal to 0.

Research Goal

Goal: *Given the specific ERGM with edge, 2-paths, and triangle counts as sufficient statistics, we want to understand the geometry of the variety of the sufficient statistics by exploring*

1. *Dimension*
2. *Irreducibility*
3. *Singularities*

Review - Background and Notation

Adjacency Matrix: *Let $G = (V, E)$ be a graph with n vertices. Let A be the adjacency matrix of the graph, i.e.*

$$A_{ij} = \begin{cases} x_{ij} & \text{if } i < j, \\ 0 & \text{if } i = j, \\ x_{ji} & \text{if } i > j. \end{cases} \tag{3}$$

Network Statistics: *The sufficient statistics of the edge-2path-triangle model are:*

- The number of edges of G: $t_1 = \sum_{i<j} A_{ij}$.
- The number of 2-paths in G: $t_2 = \sum_{i<j} (A^2)_{ij}$.
- The number of triangles in G: $t_3 = \frac{1}{6}\text{tr}A^3$.

The Variety of Sufficient Statistics

Given a graph G and network statistic $t = T(G) = (t_1, t_2, t_3)$, the reference ideal of t is:

$$I_t = \langle \sum_{i<j} x_{ij} - t_1, \\ \sum_{i<j<k} (x_{ij}x_{jk} + x_{ij}x_{ik} + x_{ik}x_{jk}) - t_2, \\ \sum_{i<j<k} x_{ij}x_{jk}x_{ik} - t_3 \rangle \\ \subseteq k[x_{ij} \mid 1 \leq i < j \leq n].$$

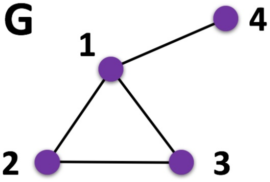
We will call the variety defined by I_t the reference variety:

$$V_t = V(I_t).$$

Example I

Consider the graph G in the figure below.

1. The number of edges of G: $t_1 = \sum_{i<j} A_{ij} = 4$.
2. The number of 2-paths in G: $t_2 = \sum_{i<j} (A^2)_{ij} = 5$.
3. The number of triangles in G: $t_3 = \frac{1}{6} tr(A^3) = 1$.



degree(I_t) = 6
dim(V_t) = 3

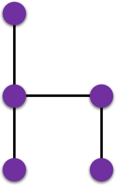
So, $t = T(G) = (4, 5, 1)$.

The reference ideal of G is:

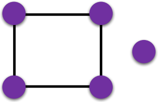
$$\begin{aligned} I_t = \langle & x_{12} + x_{13} + x_{23} + x_{14} + x_{24} + x_{34} - 4, \\ & x_{12}x_{13} + x_{12}x_{23} + x_{13}x_{23} + x_{12}x_{14} + x_{13}x_{14} + x_{14}^2 + x_{12}x_{24} + \\ & + x_{23}x_{24} + x_{24}^2 + x_{13}x_{34} + x_{23}x_{34} + x_{14}x_{34} + x_{24}x_{34} + x_{34}^2 - 5, \\ & x_{12}x_{13}x_{23} + x_{12}x_{14}x_{24} + x_{13}x_{14}x_{34} + x_{23}x_{24}x_{34} - 1 \rangle. \end{aligned}$$

Example II

$t_1 = (4, 4, 0)$, V_{t1} contains:

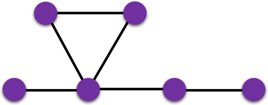


fork - chair



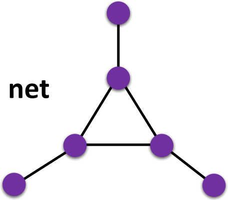
$C_4 \cup K_1$

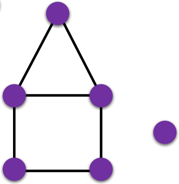
$t_2 = (6, 9, 1)$, V_{t2} contains:



X_{169}

net





co-4-fan

A Gröbner basis for I_t

Definition

Given an ideal $I \subset k[x_1, \dots, x_n]$. A finite subset G of I is a **Gröbner basis** with respect to the term order $\prec$ if the initial terms of the elements in G suffice to generate the initial ideal:

$$in_{\prec}(I) = \langle in_{\prec}(g) : g \in G \rangle.$$

Note: There is no minimality requirement for being a Gröbner basis. Hence, there is infinitely many Gröbner basis for an ideal I. But a **reduced Gröbner basis** is unique for every ideal I.

A Gröbner basis for I_t

Proposition

The polynomials

$$g_1 = f_1,$$
$$g_2 = \left(\sum_{i=2}^{n} x_{1i} + x_{2i}\right) f_1 - f_2,$$
$$g_3 = f_3 - \left(\sum_{i=3}^{n} x_{1i} x_{2i}\right) f_1 + x_{23} g_2$$

form a reduced Gröbner basis for I_t. In particular,

$$In(I_t) = \langle x_{12}, x_{13}^2, x_{23}^3 \rangle.$$

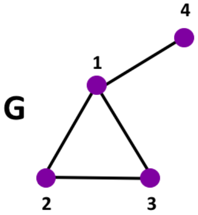
Dimension

Theorem

The ideal I_t has codimension 3.

That is, the dimension of the reference variety V_t is:

$$\dim(V_t) = \binom{n}{2} - 3.$$



$t = T(G) = (4, 5, 1)$

$$\dim(V_t) = \binom{4}{2} - 3$$
$$= 6 - 3 = 3$$

Irreducibility

Theorem

If $n \geq 4$, then I_t is prime.

In other words, if $n \geq 4$, then V_t is irreducible.

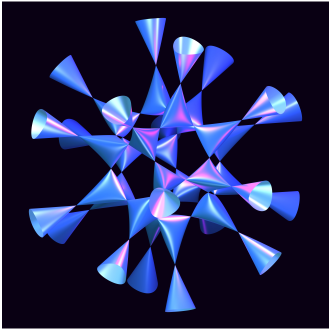
Proof.

- Fix $h_1, h_2 \notin I_t$ such that $h_1 h_2 \in I_t$. Based on the Gröbner basis, h_1 and h_2 have specific forms.
- $n = 4$: perform the polynomial long division on $h_1 h_2$ by g_2 and g_3 in Macaulay2. The result indicates that either $h_1 = 0$ or $h_2 = 0$.
- $n > 4$: by using division algorithm and induction, we show that the division of $h_1 h_2$ by g_2 and g_3 yields a remainder $r \neq 0$.

□

Definition

Intuitively, a **singular point** of $\mathbf{V}(f)$ is a point where the tangent line fails to exist.

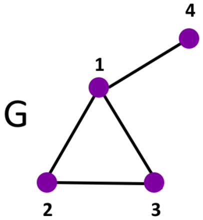


Bath Sextic - *https://imaginary.org/gallery/oliver-labs*

Edge-triangle Model:

Let G be a graph with n vertices. Let $t = T(G) = (t_1, t_3)$. The Jacobian of V_t at G, denoted $Jac(V_t)|_G$, is an $(n^2 - n) \times 2$ matrix whose rows are indexed by pairs (r, s) with $1 \leq r, s \leq n$ and $r \neq s$. The (r, s)th row of $Jac(V_t)|_G$ has the following form

$$[1 \quad \# \text{ of 2-paths between } r \text{ and } s].$$

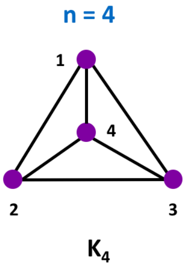


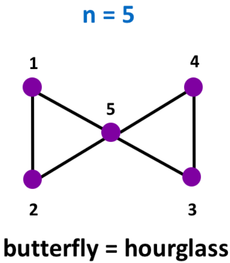
Rank 2 Jacobian

Edge-triangle Model:

Proposition

A graph G with n vertices is a singular point of a fiber of the edge-triangle model $\Leftrightarrow$ G is a windmill graph or a strongly regular graph of the form $(n, k, \frac{k^2-k}{n-1}, \frac{k^2-k}{n-1})$.





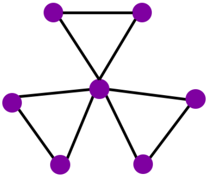
Singularitites

Remark

Just because a graph is singular in the edge-triangle model, it does not mean that it is singular in the edge-2path-triangle model.

n = 7

G



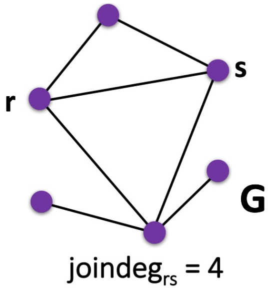
Edge-triangle Model:
Rank 1

Edge-2path-triangle Model:
Rank 3 (full rank)

Edge-2path-triangle Model:

Definition

Given a graph $G = (V, E)$ with adjacency matrix A and two vertices $r, s \in V$, let the joint degree of r and s in G be $jointdeg_{rs} := \deg r + \deg s - 2A_{rs}$.



The Jacobian of V_t at G, denoted $Jac(V_t)|_G$, is an $(n^2 - n) \times 3$ matrix whose rows are indexed by pairs (r, s) with $1 \leq r, s \leq n$ and $r \neq s$. The (r, s)th row of $Jac(V_t)|_G$ has the following form

$$[1 \quad \text{jointdeg}_{rs} \quad \# \text{ of 2-paths between } r \text{ and } s].$$

Singularities

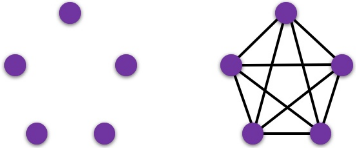
Edge-2path-triangle Model:

Proposition

Let G be a graph. If the automorphism group of G is the full symmetric group, then rank $Jac(V_t)_{|G} = 1$.

Proof.

Proof by contradiction: suppose rank $Jac(V_t)_{|G} \neq 1$.
Case 1: The number of 2-paths between r, s and u, v are different.
Case 2: The joint degrees between r, s and u, v are different.
In both cases, the permutation σ on G is not a graph automorphism (contradiction). Therefore, $Jac(V_t)_{|G} = 1$. □



Future Directions

1. Explore the relationship between singularity and degeneracy.
2. Can we use the geometry of the variety to develop a sampling algorithm for this statistical model?

References

1. M. Newman, Networks: An Introduction, *OUP Oxford* (2010).
2. D. Cox, J. Little, and D. O'Shea, Ideals, Varieties, and Algorithms, *Springer* (2000).
3. G. Chartrand, L. Lesniak, and P. Zhang, Graphs & Digraphs, Fifth Edition *CRC Press* (2010).
4. D. Eisenbud, D. Grayson, M. Stillman, and B. Sturmfels, Computations in algebraic geometry with Macaulay 2, *Springer* (2002).