Graphs over time: densification laws, shrinking diameters and possible explanations
Introduction (I)

**Key Questions**
- Evolution over time
- “Normal” Growth Patterns?

**Conventional Wisdom**
- Constant Average Degree
- Slowly growing diameter

**Recent Work**
- social, technological (routers or AS networks), scientific networks, citation graphs

**Key Parameters**
- Node degree: Heavy tailed in-degree distributions
- Node Pair Distance or Diameter (effective)
Introduction (II)

**Key Empirical Observations**

1. Densification over time (super-linearity of edges) \( e(t) \sim n(t)^a; a \in (1, 2) \)

2. Shrinking of average distance (not slow increase)

**Objective**: Graph Generation Models e.g.

(I) Community Guided Attachment – Hierarchical model

(II) “Forest Fire”

**Why is this important?**

- **Graph generation**: ‘what if’, simulations, extrapolations (real graphs impossible to collect)
- **Graph sampling**: large dataset processing needs sampling ⇒ Densification laws alleviate bad sampling - easier rejection
- **Extrapolations**: past ⇔ future, scenario validation for graph evolution
- **Abnormality detection and computer network management**: “normal” behavior ⇒ subgraphs densification law exponent \( \alpha \) ⇒ detect activity or structures that deviate (fraud, spam, or (DDoS) attacks)
Observations I: Densification Law Exponents

- 9 datasets - 4 different sources
- graph $G(t)$ up to $t$
- DPL plot: $e(t) = f(n(t))$, log-log scale

1. ArXiv Citation Graph
   - Directed $(i,j)$ in $E$ if a paper $i$ cites paper $j$
   - Time: January 1993—April 2003 – monthly snapshots
   - Densification Law Plot $\Rightarrow$ exponent $\alpha = 1.68$ ($>>1$: deviation from linear growth)
   - Out-Degree: Phantom nodes/edges phenomenon due to ‘missing past’

2. Patents Citation Graph
   - 37 years (January 1, 1963 to December 30, 1999)
   - $\alpha = 1.66$

3. Autonomous Systems Graph
   - Communication network between Autonomous Systems (AS) (sub-graphs) $\Rightarrow$ traffic between peers
   - (source: Border Gateway Protocol logs)
   - Time: 785 days (November 8 1997 to January 2 2000)
   - Both addition and deletion of nodes/edges
   - $\alpha = 1.18$

4. Affiliation Graphs
   - Bipartite graphs from ArXiv
   - edge between people to the papers authored, with co-authorship i.e. nodes at least 1 edge with same paper-node
   - Time: April 1992—March 2002
   - $\alpha \approx 1.08 – 1.15$ (5 sets)
Observations II: Shrinking Effective Diameter

**Definitions:**
- \( g(d) \): fraction of connected node pairs with shortest path < \( d \)
- hop-plot: pairs \((d, g(d))\); i.e. cdf of distances between connected node pairs (interpolation to extend to real line)

**Effective diameter** = the value of \( d \) s.t \( g(d) = 0.9 \)

Robustness: Could this result from artifacts?
(a) sampling problems: shortest paths estimation
BFS, Approximate Neighborhood Function (ANF)
(b) Disconnected components
(c) “missing” past
  - Pick \( t_0 > 0 \) (cut-off time) as beginning of data
  - put back in the nodes/edges from before time \( t_0 \)

Measurements:
(i. Full graph: effective \( d \) of the graph’s giant component
(ii. Post-\( t_0 \) sub-graph: as if we knew full past, i.e.
  for \( t > t_0 \), graph with all nodes dated before \( t \)
  \( \Rightarrow \) effective diameter of the sub-graph of nodes
dated in \([t_0, t] \)
(iii. Post-\( t_0 \) sub-graph, no past: \( \forall \) nodes dated before \( t_0 \) delete all out-links
(d) Emergence of giant component
3.2 Shrinking Diameters

The effective diameter of a graph over time is defined as the maximum distance over all connected node pairs, i.e.,
\[ d = \max_{u,v} d(u,v), \]
where \( d(u,v) \) is the shortest connecting path between two nodes \( u \) and \( v \).

The effective diameter serves a decreasing trend for all the graphs. We performed experiments to account for (a) possible sampling problems, (b) weakly connected components, (c) inherent phenomenon, and (d) intrinsic, phenomenon.

We note that the effective diameters for very large graphs, as well as a basic sampling approach in which we ran different implementations, which can estimate effective diameters to be actually close to effective diameters used in earlier work: the minimum effective diameter used in earlier work: the minimum filtration graph. Figures 1(d) and 2(d) show the filtration graphs for the Post '95 subgraph, Post '85 subgraph, no past, and full graph, respectively.

We now discuss the behavior of the effective diameter over time; one observes that it is decreasing in all graphs in our study, such as the AS dataset, it is possible for the effective diameter over time, for this collection of network datasets. Following the interesting question to be whether we could detect the decreasing effective diameter, we expected the under-estimation of the effective diameter to be actually close to the effective diameter used in earlier work: the minimum filtration graph.

Disconnected components:

Possible sampling problems:
(d) Emergence of giant component

e.g. diameter of GCC shrinks in ER model

*Could this lead to shrinking diameters?*

Answer: NO

Remarks:

• No past graph: smaller size GCC but deleted nodes negligible at t grows large
• After 4 years: 90% of nodes in GCC
• “mature” graph exhibits same behavior
Proposed Models

- **Objective**: models to capture phenomena:
  - Densification power law
  - Shrinking diameters
  - Heavy tailed in-out degree distributions
  - Small world phenomena (6 degrees of separation)

\[ e(t) \sim n(t)^a \]

- **How?**: each new node \( n(t)^{a-1} \) out-links ✗ (no insight)
- **Instead**: Self-similarity, “communities within communities”
- **Examples** of recursive patterns: geography based links in computer networks, smaller communities easier ties in social nets, hyperlinks for related topics WWW

- **Qualitative**: linkage probability \( \downarrow \) height of least common ancestor
- **Quantitative**: Difficulty Constant = measure of difficulty in crossing communities
Community Guided Attachment (CGA)

- Tree $T$, height $H$, branching factor $b$
- Number of leaves $n = |V|; n = b^H$
- $h(v,w)$: height of smallest sub-tree containing both $v,w$

**Difficulty Function (linkage probability)**
- decreasing in $h$
- scale-free i.e level independent $\frac{f(h)}{f(h-1)} = \text{const.}$
- $\Rightarrow f(h) = f(0)c^{-h}; f(0) = 1$, $c \geq 1$: difficulty const.

**THEOREM 1.** In the CGA random graph model, the expected average out-degree $d$ of a node is proportional to

$$\bar{d} = n^{1 - \log_b(c)} \quad \text{if} \quad 1 \leq c < b$$
$$= \log_b(n) \quad \text{if} \quad c = b$$
$$= \text{constant} \quad \text{if} \quad c > b$$

Note: if $c < b$, densification law with $\alpha > 1$, $\alpha = 2 - \log_b(c)$ and takes all values in $(1,2]$ as long as $c$ in $[1, b)$
Dynamic Community Guided Attachment

Idea: nodes also connecting to internal nodes of $\Gamma$

**Construction:** In time step $t$

1) add $b$ new leaves as children of each leaf: $b$-ary tree of depth $(t-1) \rightarrow$ depth $t$

2) form out-links with any node independently w.p. $c^{-d(v,w)/2}$

**Extension:**

- **tree-distance $d(v,w)$**: length of path between $v,w$ in $\Gamma$ e.g. for leaves: $d(v,w) = 2h(v,w)$

**THEOREM 2.** The dynamic CGA model has the following properties:

- When $c < b$, the average node degree is $n(1- \log_b(c))$ & in-degrees $\sim$ Zipf distribution with exponent $1/2\log_b(c)$

- When $b < c < b^2$, the average node degree is constant, and the in-degrees $\sim$ Zipf distribution with exponent $1 - 1/2\log_b(c)$

- When $c > b^2$, the average node degree is constant and the probability of an in-degree exceeding any constant bound $k$ decreases exponentially in $k$

---

**Power Law Distribution**

**Zipf's law:** Given natural language utterances: frequency of a word $\sim 1/rank$
Forest fire model

**Idea:** link \( \{v_i\} \) \( i = 1, \ldots, k \) to nodes of \( G \) s.t. \( G_{\text{final}} \):
- “rich get richer” attachment \( \to \) heavy-tailed in-degrees
- “copying” model \( \to \) communities
- Community Guided Attachment \( \to \) Densification Power Law
- shrinking effective diameters

**Construction**

i. nodes arrive in over time

ii. “center of gravity” in some part of the network

iii. linkage probability decreases rapidly with their distance from this center of gravity

**Result:** very large number of out-links:

- skewed out-degree distrib. &
- “bridges” for disparate sub-graphs
Forest fire model

**Model Formulation**

\{t=1\} \( G_1 = \{v\} \)

\{t>1\} node \( v \) joins \( G_t = \) graph up to \( t \):

(i) \( v \) first chooses an **ambassador** node \( w \) uniformly

(ii) for a r.v. \( x \) (binomially distributed), \( E(x) = 1/(1 - p) \), \( v \) selects \( x \) nodes incident to \( w \), in-links w.p. \( r \) times less than out-links (\( w_1, w_2, \ldots, w_x \))

(iii) \( v \) forms out-links to \( w_1, w_2, \ldots, w_x \), and then applies step (ii) recursively to each of \( w_i \) (nodes not re-visited \( \Rightarrow \) no cycling)

**Extensions**

*“Orphans”*: isolated nodes eg. Papers in arXiv citing non-arXiv ones \( \Rightarrow \) 2 incorporation approaches: (a) start with \( n_0 > 1 \) nodes and (b) probability \( q>0 \) no links for newcomer

**Multiple ambassadors**: newcomers choose more than one ambassador with some positive probability i.e burning links starting from 2 or more nodes

\[ p=\text{forward burning prob.} \]

\[ p_b=\text{backward burning prob.} \]

\[ r=p_b/p \]
Results

Heavy-tailed in and out-degree distributions
Phase Plot

Contour plot of densification exponent $a$ with respect to
(a) Backward Burning Ratio $r$ (y-axis)
(b) Forward Burning Probability $p$ (x-axis)

- white color: $a = 1$ (constant average degree)
- black color: $a = 2$ (graph is “dense”)
- Grey area: very narrow $\Rightarrow$ sharp transition

**Densification exponent**

**Diameter**

*Note: fit a logarithmic function $d = \alpha \log t + \beta$ ($t$ is the current time $\sim n(t)$)
Coefficient $\alpha<0$: decreasing diameter*
Conclusions

Innovation: time evolution of real graphs

Main Findings:
- Densification Power Law: average out-degree grow
- Shrinking diameters: in real networks, contrary to standard assumption, the diameter may exhibit a gradual decrease as the network grows
- Community Guided Attachment model can lead to the Densification Power Law, only one parameter suffices
- Forest Fire Model, based on only two parameters, captures patterns observed both in previous work and in the current study: heavy-tailed in- and out-degrees, the Densification Power Law, and a shrinking diameter

Potential relevance in multiple settings:
- ‘what if’ scenarios
- forecasting of future parameters of computer and social networks
- anomaly detection on monitored graphs
- designing graph sampling algorithms
- realistic graph generators
Thank you 😊

Questions???