Basic principles of algorithmic graph mining
Lecture 1: Introduction to graph mining

Aristides Gionis
Aalto University

Estonian Summer School on Computer and System Science 2016
Aug 22–24, 2016

course agenda

- introduction to graph mining
- computing basic graph statistics
- finding dense subgraphs
- spectral graph analysis
- additional topics and applications

graphs: a simple model

- entities — set of vertices
- pairwise relations among entities
  — set of edges
- can add directions, weights,…
- used to model many real-world datasets

analysis of graph datasets in the past

- graphs datasets have been studied in the past
  e.g., networks of highways, social networks
- usually these datasets were small
- visual inspection can reveal useful information

analysis of graph datasets now

- more and larger networks are collected
- networks of thousands, millions, or billions of nodes
- impossible to visualize
The internet map
types of networks
- social networks
- knowledge and information networks
- technology networks
- biological networks

social networks
- links denote a social interaction
- networks of acquaintances
- collaboration networks
  - actor networks
  - co-authorship networks
  - director networks
- phone-call networks
- e-mail networks
- IM networks
- sexual networks

knowledge and information networks
- nodes store information
  - links associate information
    - citation network (directed acyclic)
    - the web (directed)
    - peer-to-peer networks
    - word networks
    - networks of trust
    - software graphs
    - bluetooth networks
    - home page/blog networks

technological networks
- networks built for distribution of a commodity
  - the internet, power grids, telephone networks, airline networks, transportation networks

biological networks
- biological systems represented as networks
  - protein-protein interaction networks
  - gene regulation networks
  - gene co-expression networks
  - metabolic pathways
  - the food web
  - neural networks
**photo-sharing site**

what is the underlying graph?
- **nodes**: photos, tags, users, groups, albums, sets, collections, geo, query, …
- **edges**: upload, belong, tag, create, join, contact, friend, family, comment, fave, search, click, …
- tons of interesting graphs to work with
  - **tag graph**: based on photos
  - **tag graph**: based on users
  - **user graph**: based on favorites
  - **user graph**: based on groups
- which graph to pick? — depends on the application

**network science**

- the world is full with networks
- what do we do with them?
  - understand their topology and measure their properties
  - study their evolution and dynamics
  - create realistic models
  - create algorithms to make sense of network data

**properties of real-world networks**

diverse collections of graphs arising from different phenomena
— are there **typical patterns**? yes!
- **static networks**
  - heavy tails
  - clustering coefficients
  - communities
  - small diameters
- **time-evolving networks**
  - concentration
  - shrinking diameters

**heavy tails**

What do the proteins in our bodies, the Internet, a cool collection of atoms and sexual networks have in common? One man thinks he has the answer and it is going to transform the way we view the world.

Scientist 2002

Albert-László Barabási
**degree distribution**

- \( C_k \) = number of vertices with degree \( k \)

- problem: find the probability distribution that fits best the observed data

**power-law degree distribution**

- \( C_k \) = number of vertices with degree \( k \), then
  \[ C_k = c k^{-\gamma} \]
  with \( \gamma > 1 \), or
  \[ \ln C_k = \ln c - \gamma \ln k \]

- plotting \( \ln C_k \) versus \( \ln k \) gives a straight line with slope \(-\gamma\)

- heavy-tail distribution: there is a non-negligible fraction of nodes that has very high degree (hubs)

- scale free: average is not informative

---

**power-law degree distribution**

Power-laws in a wide variety of networks [Newman, 2003]

Sheer contrast with Erdős-Rényi random graphs

**power-law degree distribution**

Do the degrees follow a power-law distribution?

Three problems with the initial studies

- Graphs generated with traceroute sampling, which produces power-law distributions, even for regular graphs [Lakhina et al., 2003].

- Methodological flaws in determining the exponent see [Clauset et al., 2009] for a proper methodology.

- Other distributions could potentially fit the data better but were not considered, for example, lognormal.

Disclaimer: we will be referring to these distributions as heavy-tailed, avoiding a specific characterization.

**power-law degree distribution**

All networks above have the same degree sequence but structurally are very different [Li et al., 2005]

**maximum degree**

- For random graphs, the maximum degree is highly concentrated around the average degree \( z \)

- For power-law graphs
  \[ d_{\text{max}} \approx n^{1/(\gamma-1)} \]

- Hand-waving argument: solve \( n \Pr[X \geq d] = \Theta(1) \)
**heavy tails, eigenvalues**

- log-log plot of eigenvalues of the Internet graph in decreasing order
- again a power law emerges [Faloutsos et al., 1999]

**heavy tails, triangles**

- triangle distribution in flickr
- figure shows the count of nodes with $k$-triangles vs. $k$ in log-log scale
- again, heavy tails emerge [Tsourakakis, 2008]

**clustering coefficients**

- a proposed measure to capture local clustering is graph transitivity
  \[ T(G) = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \]
- captures "transitivity of clustering"
- if $u$ is connected to $v$ and $v$ is connected to $w$, it is also likely that $u$ is connected to $w$

**community structure**

loose definition of community: a set of vertices densely connected to each other and sparsely connected to the rest of the graph

**community structure**

[Leskovec et al., 2009]

- study community structure in an extensive collection of real-world networks
- introduce the network community profile (NCP) plot
- characterizes the best possible community over a range of scales

**community structure**

[Leskovec et al., 2009]

- dolphins network and its NCP
**community structure**

- do large real networks have such nice structure?  **NO!**

![Local Spectral vs Metis+MQI](image)

NCP of a DBLP graph (source [Leskovec et al., 2009])

**important findings of [Leskovec et al., 2009]**

1. up to certain size \( k \) (\( \approx 100 \) vertices) there are good cuts
   - as the size increases so does the quality of the community
2. at the size \( k \) we observe the best possible community
   - such communities are typically connected to the remainder with a single edge
3. above the size \( k \) the community quality decreases
   - this is because they blend in and gradually disappear

**small-world phenomena**

*small worlds*: graphs with short paths

- Stanley Milgram (1933-1984)  
  “The man who shocked the world”
- obedience to authority (1963)
- small-World experiment (1967)
  — we live in a small-world

for criticism on the small-world experiment, see

*Could It Be a Big World After All? What the Milgram Papers in the Yale Archives Reveal About the Original Small World Study* by Judith Kleinfield

**small-world experiments**

- letters were handed out to people in Nebraska to be sent to a target in Boston
- people were instructed to pass on the letters to someone they knew on first-name basis
- the letters that reached the destination (64 / 296) followed paths of length around 6
  - *Six degrees of separation* : (play of John Guare)
  - also:
    - the Kevin Bacon game
    - the Erdős number

**small diameter**

proposed measures

- **diameter**: largest shortest-path over all pairs
- **effective diameter**: upper bound of the shortest path of 90% of the pairs of vertices
- **average shortest path**: average of the shortest paths over all pairs of vertices
- **characteristic path length**: median of the shortest paths over all pairs of vertices
- **hop-plots**: plot of \( |N(u)| \), the number of neighbors of \( u \) at distance at most \( h \), as a function of \( h \)
  [Faloutsos et al., 1999].

**time-evolving networks**

- **densification power law**:  
  \[ |E_t| \propto |V_t|^\alpha \quad 1 < \alpha \leq 2 \]
- **shrinking diameters**: diameter is shrinking over time.

*J. Leskovec*  
*J. Kleinberg*  
*C. Faloutsos*

[Leskovec et al., 2005]
random graphs

- a random graph is a set of graphs together with a probability distribution on that set
- example

```
1   2
\quad Probability \frac{1}{3}
3
```

a random graph on \{1, 2, 3\} with 2 edges with the uniform distribution

random graphs

- Erdős-Rényi (or Gilbert-Erdős-Rényi) random graph model
- the \( G(n, p) \) model:
  - \( n \): the number of vertices
  - \( 0 \leq p \leq 1 \): probability
  - for each pair \( (u, v) \), independently generate the edge \( (u, v) \) with probability \( p \)
  - \( G(n, p) \) a family of graphs, in which a graph with \( m \) edges appears with probability \( p^m (1 - p)^{n-m} \)
- the \( G(n, m) \) model: related, but not identical

properties of random graphs

- a property \( P \) holds almost surely/with high probability (whp \( \rightarrow 1 - \Theta(1) \)) if
  \[
  \lim_{n \to \infty} \Pr[ G \text{ has } P ] = 1
  \]
- which properties hold as \( p \) increases?

properties of random graphs

- a property \( P \) holds almost surely/with high probability (whp \( \rightarrow 1 - \Theta(1) \)) if
  \[
  \lim_{n \to \infty} \Pr[ G \text{ has } P ] = 1
  \]
- which properties hold as \( p \) increases?
- threshold phenomena: many properties appear suddenly
  - there exist a probability \( p_c \) such that
    - for \( p < p_c \) the property does not hold a.s.
    - for \( p > p_c \) the property holds a.s.
**the giant component**

- let $z = np$ be the average degree
- if $z < 1$ the largest component has size $O(\log n)$ a.s.
- if $z > 1$ the largest component has size $\Theta(n)$ a.s.;
  the second largest component has size $O(\log n)$ a.s.
- if $z = \omega(\log n)$ the graph is connected a.s.

**phase transition**

- if $z = 1$ there is a phase transition
  - the largest component has size $O(n^{2/3})$
  - the sizes of the components follow a power-law

**degree distribution**

- degree distribution: binomial
  \[ C_k = \binom{n-1}{k} p^k (1-p)^{n-k-1} \]
- the limit distribution of the normalized binomial distribution
  $\text{Bin}(n, p)$ is the normal distribution provided that
  $np(1-p) \to +\infty$ as $n \to +\infty$
- if $p = \frac{3}{2}$ the limit distribution of $\text{Bin}(n, p)$ is the
  Poisson distribution

**random graphs and real datasets**

- a beautiful and elegant theory studied exhaustively
- have been used as idealized generative models
- unfortunately, they don’t always capture reality…

**models**

- growth with preferential attachment
- structure $\rightarrow$ randomness $\rightarrow$ small-world networks
- forest-fire model

**models of real-world networks**
preferential attachment

R. Albert  L. Barabási  B. Bollobás  O. Riordan

growth model:
- at time \( n \), vertex \( n \) is added to the graph
- one edge is attached to the new vertex
- the other vertex is selected at random with probability proportional to its degree
- obtain a sequence of graphs \( \{ G_{n}(\theta) \} \)
- power law distribution arises!

small-world models

Duncan Watts  Steven Strogatz

construct a network with
- small diameter
- positive density of triangles

small-world models

Watts-Strogatz graph on 4 000 vertices, starting from a 10-regular graph
- intuition: if you add a little bit of randomness to a structured graph, you get the small world effect
- related work: see [Bollobás and Chung, 1988]

small-world models

Watts-Strogatz on 1 000 vertices with rewiring probability \( p = 0.05 \)

navigation in a small world

Jon Kleinberg

how to find short paths using only local information?
- we will use a simple directed model [Kleinberg, 2000]
- a local algorithm
  - can remember the source, the destination and its current location
  - can query the graph to find the long-distance edge at the current location
**navigation in a small world**

\[ d(u, v) \]: shortest path distance using only original grid edges

directed graph model, parameter \( r \):

- each vertex is connected to its four adjacent vertices
- for each vertex \( v \) we add an extra link \((v, u)\) where \( u \) is chosen with probability proportional to \( d(v, u) \)

**notice**: compared to the Watts-Strogatz model the long range edges are added in a **biased** way

(source [Kleinberg, 2000])

---

**forest-fire model**

- \( r = 0 \): random edges, independent of distance
- as \( r \) **increases** the length of the long distance edges **decreases** in expectation

**results**

1. \( r < 2 \): the end points of the long distance edges tend to be uniformly distributed over the vertices of the grid
   - is unlikely on a short path to encounter a long distance edge whose end point is close to the destination
   - no local algorithm can find them
2. \( r = 2 \): there are **short paths**
   - a short path can be found be the simple algorithm that always selects the edge that takes closest to the destination
3. \( r > 2 \): there are **no short paths**, with high probability

[Leskovec et al., 2007] proposed the forest-fire model that is able to re-produce at a qualitative scale most of the established properties of real-world networks

---

**acknowledgements**

Paolo Boldi  Charalampos Tsourakakis
references

The diameter of a cycle plus a random matching.

Power-law distributions in empirical data.

On power-law relationships of the internet topology.
In *SIGCOMM*.

Navigation in a small world.

Sampling biases in ip topology measurements.

references (cont.)

Graphs over time: densification laws, shrinking diameters and possible explanations.

Graph evolution: Densification and shrinking diameters.
*ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2.

Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters.

Towards a theory of scale-free graphs: Definition, properties, and implications.

references (cont.)

The structure and function of complex networks.

Fast counting of triangles in large real networks without counting: Algorithms and laws.
In *ICDM*. 