Mining Large Quasi-Cliques with Quality Guarantees from Vertex Neighborhoods

Aritra Konar
and
Nicholas D. Sidiropoulos

Department of Electrical and Computer Engineering

KDD 2020 Research Track
Dense Subgraph Discovery

- **Problem:** Given a graph, find list of “dense” subgraphs
  - A key primitive in graph mining

- **Applications:**
  - Detecting correlated genes [Tsourakakis et al. 2013]
  - Anomaly detection in e-commerce and social networks [Hooi et al. 2016]
  - Story identification in Twitter streams [Angel et al. 2012]
What is a dense subgraph?

- **Archetype:** Cliques
  - NP-hard, restrictive definition

- **Other notions:** Quasi-cliques
  - Core decomposition [Seidman 1983]
  - Average Degree [Goldberg 1984], k-Clique Densest Subgraph [Tsourakakis 2015]
  - Optimal Quasi-clique [Tsourakakis et al. 2013]

- **Algorithms:**
  - Maximum-flow [Goldberg 1984, Tsourakakis 2015, Mitzenmacher et al. 2015]
  - Semidefinite Relaxation [Cadena et al. 2016]
  - Greedy [Charikar 2000, Batagelj-Zaversnik 2003, Tsourakakis et al. 2013, Tsourakakis 2015]
  - Local-search [Tsourakakis et al. 2013]
Our approach

- Look at vertex neighborhoods!
  - List all triangles in graph [Schank 2005, Lapaty 2008, Suri-Vassilvitskii 2011]
  - Compute the local clustering coefficient (LCC) of each vertex
    - $LCC = \text{edge density of one-hop neighborhood of } v$
  - Output neighborhood with highest LCC

- But why do this?
Obtained a list of non-trivial (maximal) cliques and quasi-cliques without using any specialized methods!
- Comparison with triangle-densest subgraph [Tsourakakis 2015, Mitzenmacher et al. 2015]
  - Best neighborhood consistently outperforms dedicated algorithm!

| Graph           | Max-Flow | Neighborhood |
|-----------------|----------|--------------|
|                 | $|S|$  | $\delta(S)$ | $\tau(S)$ | $|S|$ | $\delta(S)$ | $\tau(S)$ |
| arXiv-HepPh     | 239    | 1           | 1         | 239 | 1           | 1 |
| arXiv-AstroPh   | 76     | 0.80        | 0.59      | 57  | 1           | 1 |
| arXiv-CondMat   | 30     | 0.93        | 0.72      | 23  | 1           | 1 |
| arXiv           | 146    | 0.49        | 0.25      | 74  | 1           | 1 |
| DBLP            | 114    | 1           | 1         | 114 | 1           | 1 |
| Facebook-A      | 195    | 0.79        | 0.54      | 50  | 0.94        | 0.85 |
| blogCatalog3    | 621    | 0.31        | 0.05      | 12  | 0.95        | 0.87 |
| Facebook-B      | 198    | 0.36        | 0.08      | 20  | 0.95        | 0.85 |
| loc-Gowalla     | 311    | 0.27        | 0.04      | 36  | 0.94        | 0.85 |
| web-Stanford    | 684    | 0.17        | 0.02      | 53  | 1           | 1 |
| web-Google      | 66     | 0.85        | 0.64      | 54  | 0.93        | 0.84 |
| ppi-Human       | 361    | 0.42        | 0.14      | 81  | 0.93        | 0.89 |
| email-Enron     | 388    | 0.19        | 0.02      | 14  | 0.93        | 0.82 |
| router-Caida    | 75     | 0.55        | 0.20      | 12  | 0.92        | 0.94 |
| Amazon          | 50     | 0.19        | 0.02      | 7   | 1           | 1 |

size $|S|$, edge-density $\delta(S)$, and triangle-density $\tau(S)$
Why does this happen?

- **Observation:**
  - Recurring traits of real-world graphs:
    - High clustering coefficients [Watts-Strogatz 98]
    - Power-law degree distributions [Faloutsos (x3) 99, Barabasi-Albert 99]

- **Main question:**
  - Do these properties imply that vertex neighborhoods harbor dense subgraphs of non-trivial sizes?
A note on clustering coefficients

- **Global clustering coefficient (GCC):**
  - The probability that a path of length 2 has its endpoints closed

  \[
  C_g = \frac{3(\text{# triangles in } G)}{\text{(\# paths of length 2 in } G)}
  \]

- **Useful Result:** [Gleich-Seshadhri 12]

  - Define probability distribution on vertices

    \[
    p_v = \frac{\text{(\# paths of length 2 centered at } v)}{\text{(\# paths of length 2 in } G)}, \forall \ v \in V
    \]

  - Then, \( E_p[C_v] = C_g \)
A note on clustering coefficients

- **Recall:**
  - LCC = edge density of one-hop neighborhood $\delta(\mathcal{N}_v)$

- **Corollary 1:**
  - $\mathbb{E}_p[\delta(\mathcal{N}_v)] = C_g$
  - Since $\Pr\{\delta(\mathcal{N}_v) \geq C_g\} > 0$, high GCC implies the existence of a vertex neighborhood with high edge-density

- **Corollary 2:**
  - $\text{Var}[\delta(\mathcal{N}_v)] \leq C_g(1 - C_g)$
  - High GCC implies presence of many vertex neighborhoods with high edge-density
A note on clustering coefficients

- Limitation:
  - High edge-density is necessary, but not sufficient for a neighborhood to be dense and of non-trivial size

- Counter-example:

  ![Diagram of neighborhoods](image)

  - Although $C_g = 1$, every neighborhood is simply an edge
Vertex neighborhoods as dense subgraphs

- **Desiderata:** Want to show existence of vertex neighborhood with
  - “High” edge-density
  - “large” size (degree)

- **Approach:** Invoke the probabilistic method [Alon-Spencer 16]
  - Define pair of “bad” events
    - (A) vertex sampled with probability $p_v$ has a neighborhood with “low” edge-density
    - (B) vertex sampled with probability $p_v$ has a “small” degree

- Suffices to show

$$\Pr\{A \cup B\} < 1 \Rightarrow \Pr\{A^c \cap B^c\} > 0$$
Vertex neighborhoods as dense subgraphs

- **Assumptions:**
  - (A): $\mathcal{G}$ obeys a power-law distribution with exponent 2
  - (B): $\mathcal{G}$ has no missing degrees

- **Main theorem:**
  - For every choice of $\beta \in \left(\frac{d_{\min}}{d_{\max}}, C_g\right)$
    - there exists a vertex neighborhood of size $|\mathcal{N}_v| \geq \beta d_{\max}$
    - and edge-density $\delta(\mathcal{N}_v) \geq \frac{C_g - \beta}{1 - \beta}$

- **Take-away:** high GCC and power-law distributions imply the presence of dense neighborhood subgraphs
Vertex neighborhoods as dense subgraphs

- Illustration: Facebook graph
## Experiments

### Datasets:

| Graph          | n   | m     | $d_{\text{max}}$ | $C_q$ | $\bar{C}$ |
|----------------|-----|-------|-------------------|-------|----------|
| arXiv-HepPh    | 12,008 | 112K   | 491               | 0.659 | 0.612    |
| arXiv-AstroPh  | 18,772 | 198K   | 504               | 0.318 | 0.677    |
| arXiv-CondMat  | 23,133 | 93,497 | 279               | 0.264 | 0.633    |
| arXiv          | 86,376 | 517K   | 1,253             | 0.560 | 0.678    |
| DBLP           | 317K  | 1.05M  | 343               | 0.306 | 0.632    |
| Facebook-A     | 4,039 | 88,234 | 1,045             | 0.519 | 0.605    |
| blogCatalog3   | 10,312 | 333K   | 3,992             | 0.091 | 0.463    |
| Facebook-B     | 63,731 | 817K   | 1,098             | 0.148 | 0.221    |
| loc-Gowalla    | 196K  | 950K   | 14,730            | 0.023 | 0.237    |
| Flickr         | 513K  | 3.19M  | 4,369             | 0.159 | 0.168    |
| web-Stanford   | 281K  | 2.31M  | 38,625            | 0.008 | 0.598    |
| web-Google     | 875K  | 5.10M  | 6,332             | 0.055 | 0.514    |
| PPI-Human      | 21,557 | 342K   | 2,130             | 0.119 | 0.207    |
| email-Enron    | 36,692 | 183K   | 1,383             | 0.085 | 0.497    |
| router-Caida   | 192K  | 609K   | 1,071             | 0.061 | 0.157    |
| Amazon         | 334K  | 923K   | 549               | 0.205 | 0.397    |

### What happens when GCC is small?
Experiments

- Best neighborhood can still outperform a dedicated algorithm!

--- Clique returned by GreedyOQC
[Tsourakakis et al. 2013]
--- Max. degree
--- GCC
Experiments

- Use neighborhoods as seed sets for local search [Tsourakakis et al. 2013]

--- Clique returned by GreedyOQC [Tsourakakis et al. 2013]
--- Max. degree
--- GCC
Comparing quality of seeds

| Graph            | Core decomposition | Vertex neighborhoods |
|------------------|-------------------|----------------------|
|                  | $|S|$    | $\delta(S)$ | Avg. degree | $|S|$    | $\delta(S)$ | Edge density |
| arXiv-AstroPh    | 57      | 1           | 81          | 0.75     | 57          | 1           |
| arXiv            | 146     | 0.49        | 147         | 0.52     | 75          | 0.95        |
| blogCatalog3     | 447     | 0.4         | 1550        | 0.08     | 12         | 0.95        |
| Facebook-B       | 699     | 0.12        | 723         | 0.07     | 20         | 0.95        |
| loc-Gowalla      | 183     | 0.41        | 162         | 0.27     | 36         | 0.94        |
| web-Stanford     | 387     | 0.29        | 694         | 0.17     | 71         | 0.95        |
| router-Caida     | 92      | 0.45        | 91          | 0.31     | 12         | 0.92        |
| Amazon           | 497     | 0.013       | 47          | 0.20     | 7          | 0.95        |

- Vertex neighborhoods are good seeds: Consistently yield seeds of considerably higher quality.
# Results: cliques

| Graph          | Cliques | | | |
|----------------|---------|---|---|---|
|                | NB      | NB + LS | GreedyOQC |
| **arXiv-HepPh**| 239     | 239     | 239     |
| **arXiv-AstroPh** | 57     | 57     | 57     |
| **arXiv-CondMat** | 23     | 26     | 26     |
| **arXiv**      | 74      | 74     | 74     |
| **DBLP**       | 114     | 114     | 114     |
| **Facebook-A** | 11      | 32     | 69     |
| **blogCatalog3** | 10     | 29     | 31     |
| **Facebook-B** | 12      | 25     | 25     |
| **loc-Gowalla** | 15     | 28     | 16     |
| **web-Stanford** | 53     | 53     | 14     |
| **web-Google** | 25      | 43     | 44     |
| **PPI-Human**  | 81      | 130     | 130     |
| **email-Enron** | 10     | 16     | 16     |
| **router-Caida** | 9      | 15     | 6      |
| **Amazon**     | 7       | 7       | 5      |

- **Neighborhoods are dense subgraphs:** Largest neighborhood cliques are no smaller than those computed by baselines on 6/15 datasets.
- **Vertex neighborhoods are good seeds:** Local search + proper seeds produce can produce cliques of non-trivial sizes; competitive with greedyOQC.
Results: quasi-cliques

| Graph            | $|S|$ | $\delta(S)$ | $\tau(S)$ |
|------------------|-----|-------------|------------|
|                  | NB  | NB + LS | Greedy | NB  | NB + LS | Greedy | NB  | NB + LS | Greedy |
| arXiv-HepPh      | 246 | 247     | -      | 0.95 | 0.95    | -      | 0.92 | 0.91    | -      |
| arXiv-AstroPh    | 48  | 45      | -      | 0.90 | 0.99    | -      | 0.83 | 0.97    | -      |
| arXiv-CondMat    | 19  | 18      | -      | 0.86 | 0.96    | -      | 0.68 | 0.89    | -      |
| arXiv            | 75  | 60      | -      | 0.95 | 0.98    | -      | 0.92 | 0.94    | -      |
| DBLP             | 105 | -       | -      | 0.95 | -       | -      | 0.92 | -       | -      |
| Facebook-A       | 50  | 53      | 118    | 0.94 | 0.98    | 0.97   | 0.85 | 0.94    | 0.92   |
| blogCatalog3     | 12  | 52      | 52     | 0.95 | 0.96    | 0.96   | 0.87 | 0.88    | 0.88   |
| Facebook-B       | 20  | 17      | 36     | 0.95 | 0.98    | 0.96   | 0.85 | 0.95    | 0.89   |
| loc-Gowalla      | 36  | 32      | 23     | 0.94 | 0.99    | 0.95   | 0.85 | 0.97    | 0.86   |
| web-Stanford     | 71  | 68      | 16     | 0.95 | 0.99    | 0.96   | 0.89 | 0.97    | 0.88   |
| web-Google       | 54  | 48      | 48     | 0.93 | 0.99    | 0.99   | 0.84 | 0.98    | 0.98   |
| ppi-Human        | 81  | -       | -      | 0.93 | -       | -      | 0.89 | -       | -      |
| email-Enron      | 14  | 12      | 22     | 0.93 | 0.98    | 0.96   | 0.82 | 0.95    | 0.89   |
| router-Caida     | 12  | 15      | -      | 0.92 | 0.97    | -      | 0.94 | 0.99    | 0.95   |
| Amazon           | 7   | 8       | 7      | 0.95 | 0.96    | 0.90   | 0.86 | 0.90    | 0.72   |

- **Neighborhoods are dense subgraphs**: Best neighborhood quasi-cliques are competitive in general.
- **Vertex neighborhoods are good seeds**: Yield smaller quasi-cliques with higher triangle density compared to greedy.
- **Greedy can fail to capture spectrum of subgraphs**
Conclusions

- **Neighborhoods are dense subgraphs:**
  - High clustering coefficients and power-law degree distributions imply that graphs harbor dense neighborhoods
  - In practice:
    - Neighborhoods can form large maximal cliques and quasi-cliques
    - Can serve as good seeds for local search
    - Combined approach yields state-of-the-art results
  - Simple methods work very well!

- **Future Work:**
  - Additional theoretical analysis
  - Extensions to weighted, bipartite, time-evolving networks?
Thank you!