Fraud Detection by Dense Subgraph Detection

Tong Zhao

Fraud Detection

- Text-based methods
- Graph-based methods
  - Unexpected spectral patterns

    [1] Prakash, B. Aditya, et al. "Eigenspokes: Surprising patterns and scalable community chipping in large graphs." *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*. IEEE, 2009.
  - Dense subgraphs

    [2] Hooi, Bryan, et al. "Fraudar: Bounding graph fraud in the face of camouflage." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

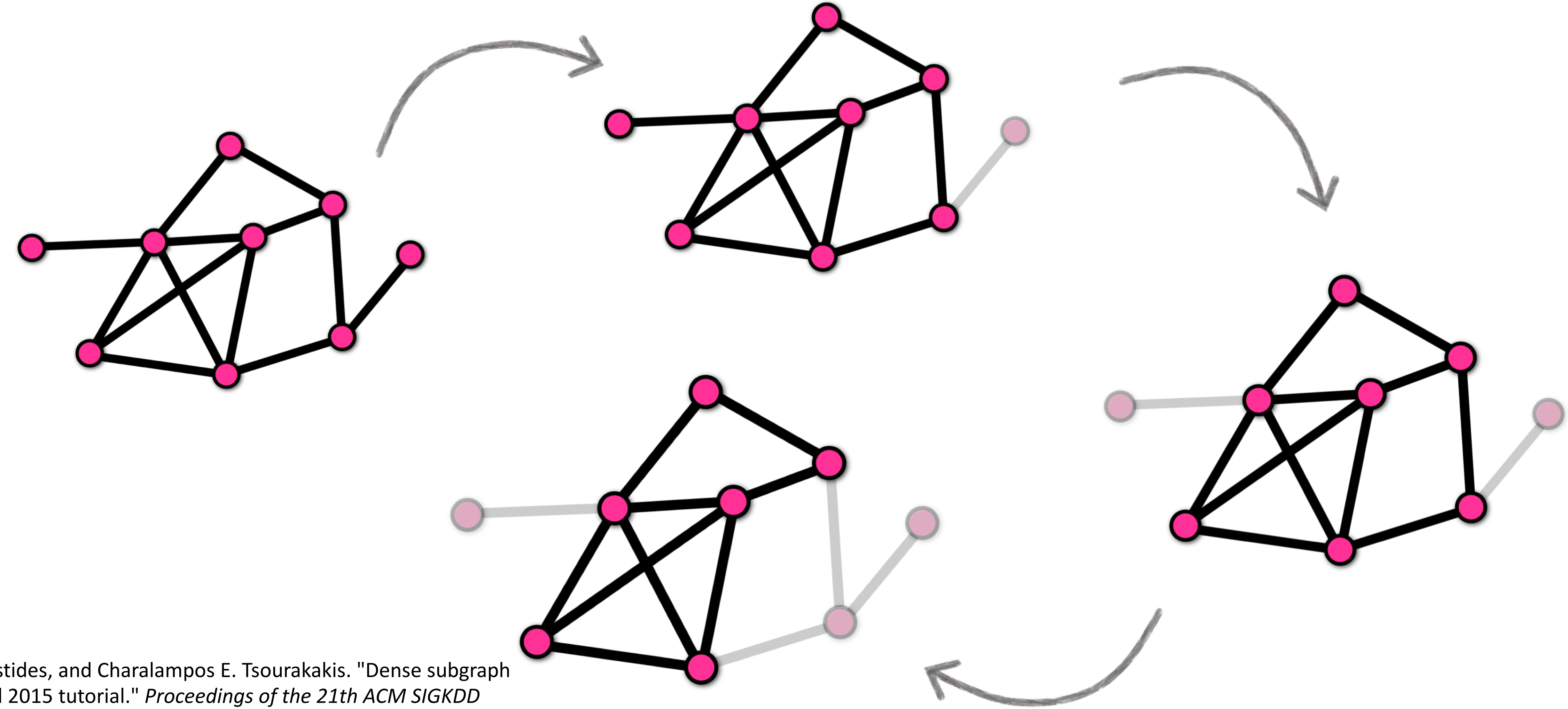
Dense Subgraph Detection

- Given a graph $G = (V, E)$ with vertices V and edges $E \subseteq V \times V$.
- Find a subgraph S such that $d(S)$ is maximized.
- GoldBerg’s algorithm (1984)
  - Transferred to a min-cut problem which can be solved as a max-flow problem.

  [3] Goldberg, Andrew V. *Finding a maximum density subgraph*. Berkeley, CA: University of California, 1984.
- Charikar’s algorithm (2000)
  - Approximation algorithm by greedy approach.
  - Provable 2-approximation guarantee.

  [4] Charikar, Moses. "Greedy approximation algorithms for finding dense components in a graph." *International Workshop on Approximation Algorithms for Combinatorial Optimization*. Springer, Berlin, Heidelberg, 2000.

Charikar’s algorithm on Undirected Graph



Figures from:
[5] Gionis, Aristides, and Charalampos E. Tsourakakis. "Dense subgraph discovery: Kdd 2015 tutorial." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

Charikar’s algorithm on Undirected Graph

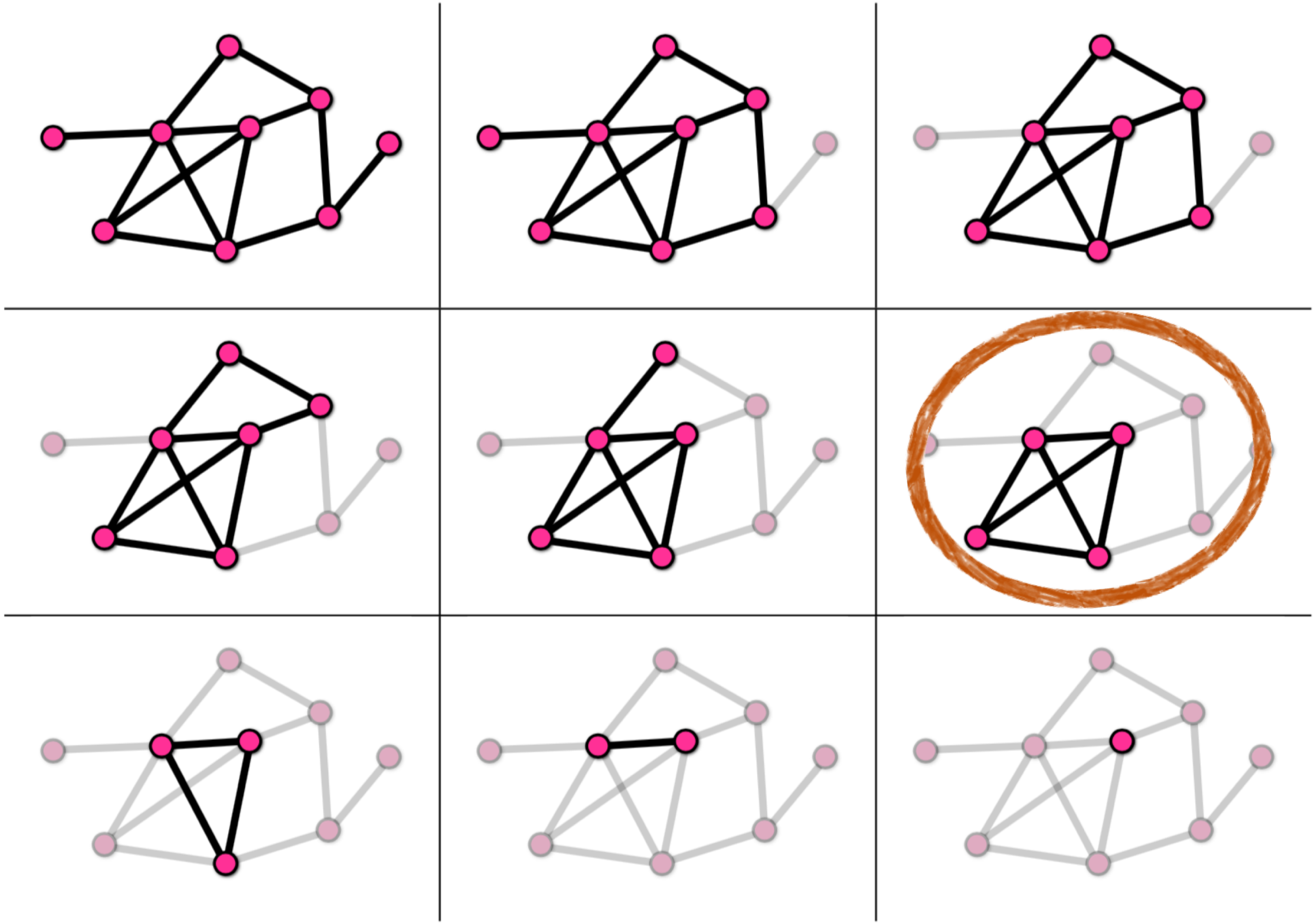


Figure from:
[5] Gionis, Aristides, and Charalampos E. Tsourakakis. "Dense subgraph discovery: Kdd 2015 tutorial." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

Charikar's algorithm on Undirected Graph

input: undirected graph $G = (V, E)$

output: S, a dense subgraph of G

1 set $G_n \leftarrow G$

2 for $k \leftarrow n$ downto 1

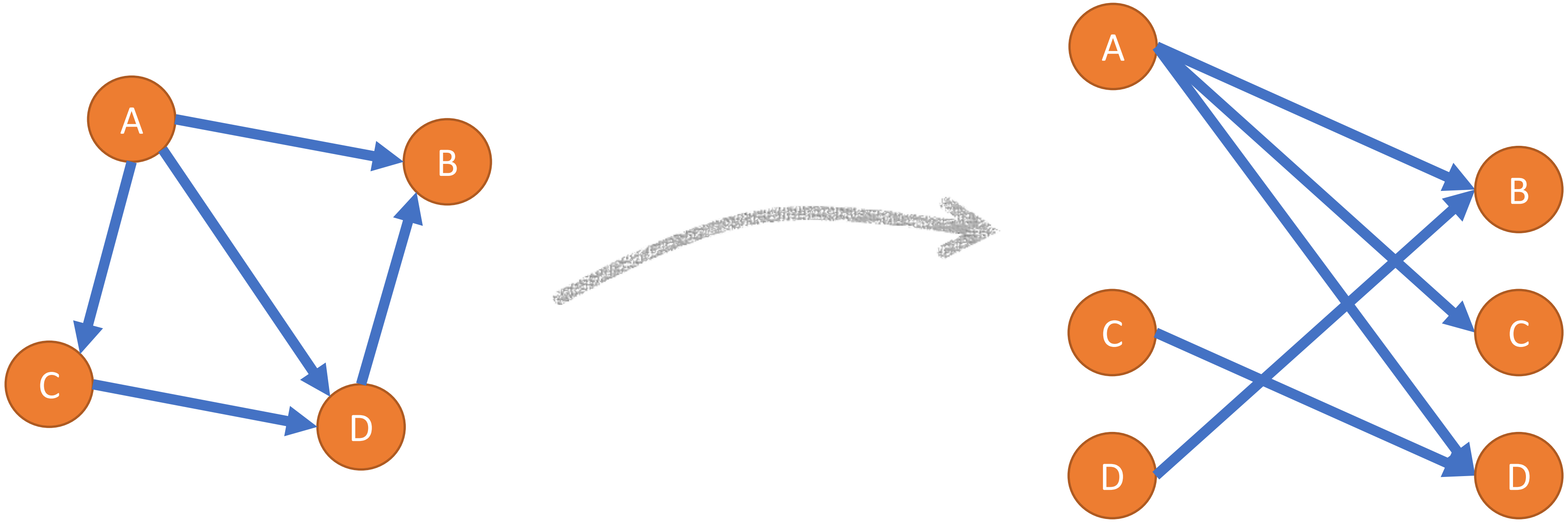
2.1 let v be the smallest degree vertex in G_k

2.2 $G_{k-1} \leftarrow G_k \setminus \{v\}$

3 output the densest subgraph among $G_n, G_{n-1}, \ldots, G_1$

On Directed Graph

- Duplicate the vertices.
- Make G into a bipartite graph.



Fraudar

Require: Bipartite $G = (\mathcal{U} \cup \mathcal{W}, \mathcal{E})$; density metric g of the form in (1)

procedure FRAUDAR (G, g)

  Construct priority tree T from $\mathcal{U} \cup \mathcal{W}$

  $\mathcal{X}_0 \leftarrow \mathcal{U} \cup \mathcal{W}$

  for $t = 1, \ldots, (m+n)$ **do**

    $i^* \leftarrow \arg\max_{i \in \mathcal{X}_i} g(\mathcal{X}_i \setminus \{i\})$

    Update priorities in T for all neighbors of i^*

    $\mathcal{X}_t \leftarrow \mathcal{X}_{t-1} \setminus \{i^*\}$

  end for

  return $\arg\max_{\mathcal{X}_i \in \{\mathcal{X}_0, \ldots, \mathcal{X}_{m+n}\}} g(\mathcal{X}_i)$

end procedure

[2] Hooi, Bryan, et al. "Fraudar: Bounding graph fraud in the face of camouflage." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

Runtime

- Utilized a priority tree
  - Fast minimum retrieval ($O(\log|V|)$)
  - Fast updating ($O(\log|V|)$)

- Total runtime: $O((|V| + |E|)\log|V|)$

Dataset

- Twitter data extracted in July 2009.
- 41.7 million users.
- 1.47 billion follows.
- Fraudar detected a 4031 followers by 4313 followees subgraph with density 68%.
- Human labeling: 57% of the detected followers are were labelled as fraudulent, deleted or suspended accounts.