file:///C:/Users/pskang/Downloads/ssf-170830042806.pdf

Graph-based Extractive Summarization

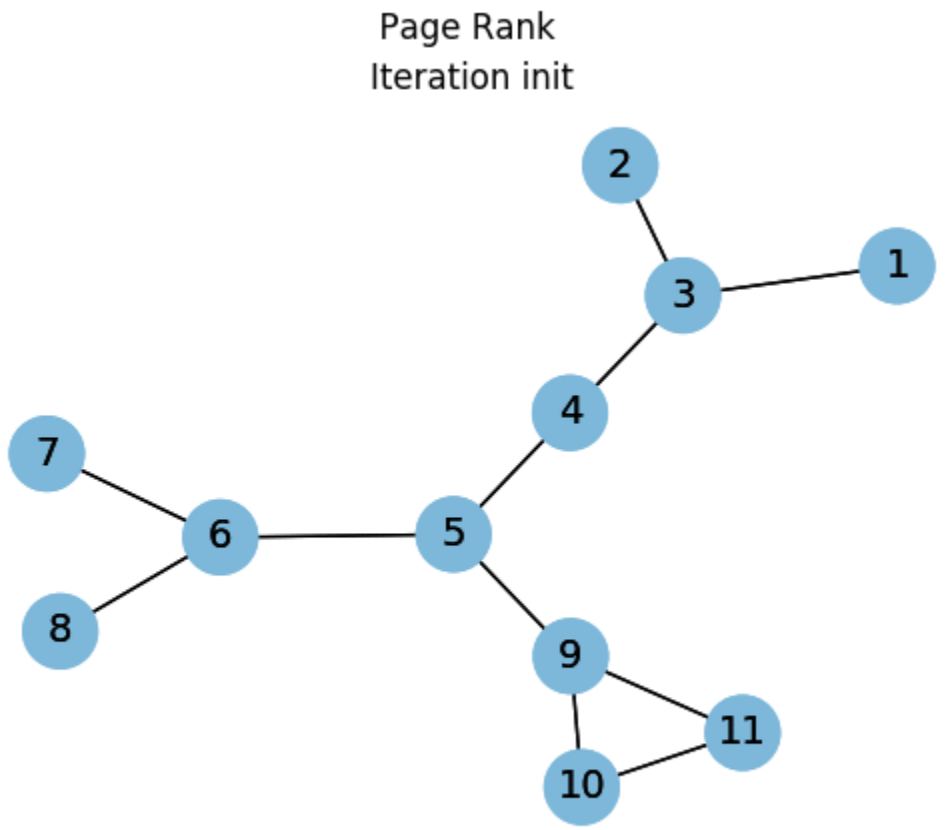
Pilsung Kang

School of Industrial Management Engineering

Korea University

Graph-based Ranking

- **Graph-based ranking**
  - ✓ A way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph.
  - ✓ Basic idea: "voting" or "recommendation"
    - One vertex links to another one, it is casting a vote for that other vertex.
    - The importance of the vertex casting the vote determines how important the vote itself is.



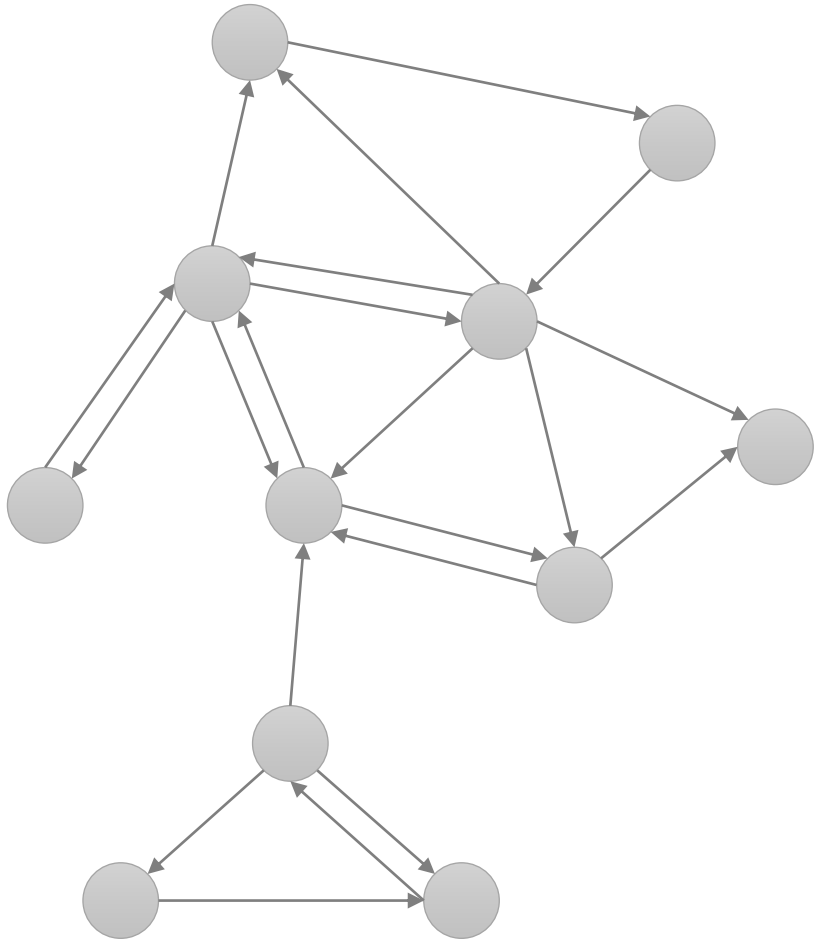
https://stellasia.github.io/blog/2020-03-07-page-rank-animation-with-networkx-numpy-and-matplotlib/

Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)

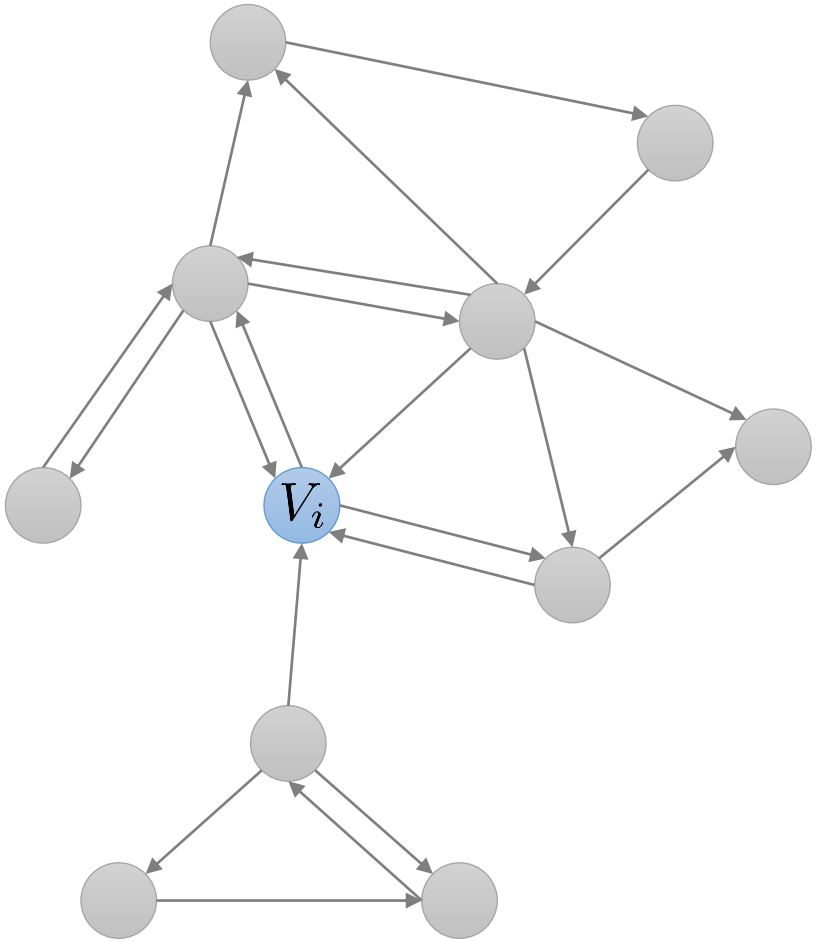


Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)

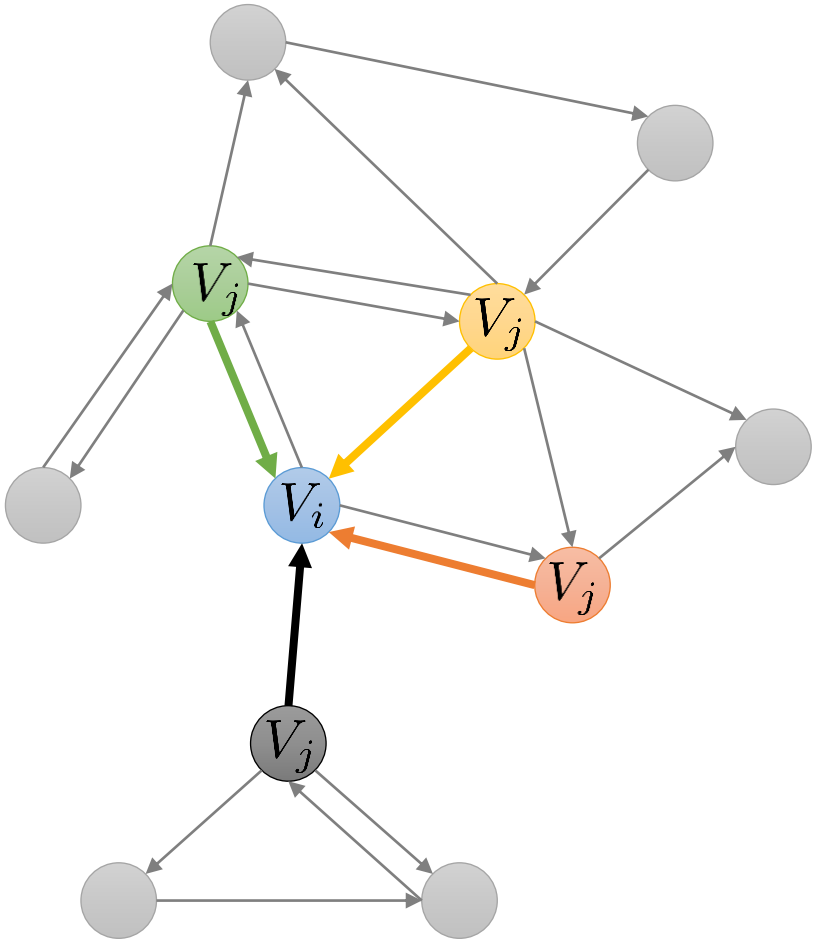


Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)

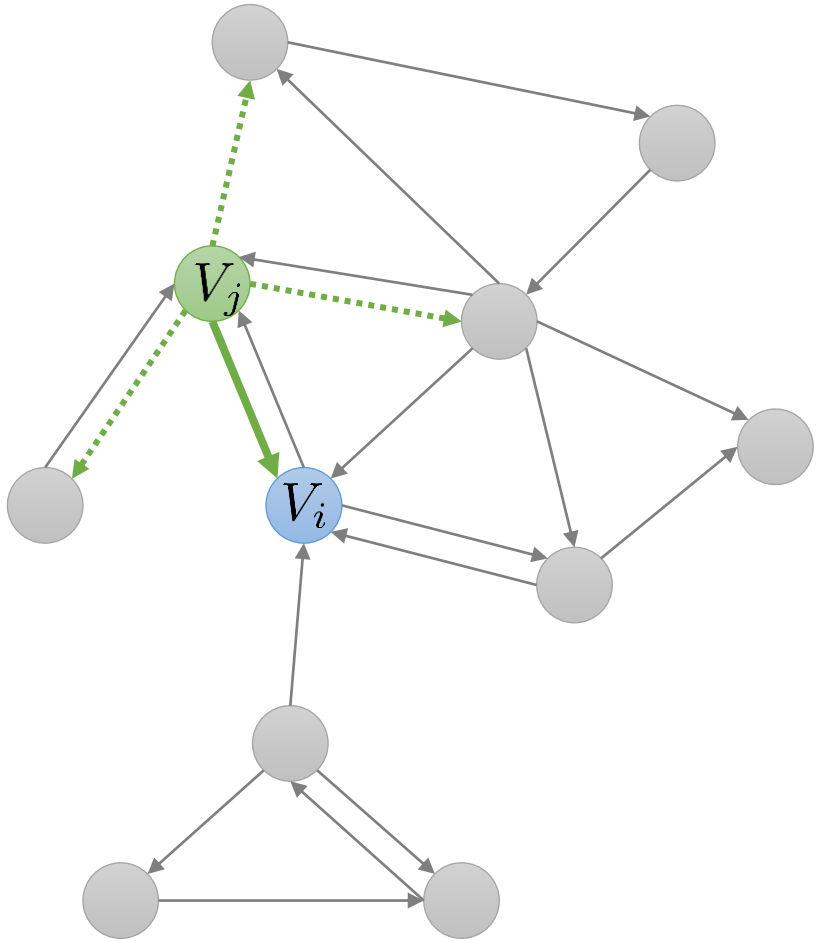


Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)

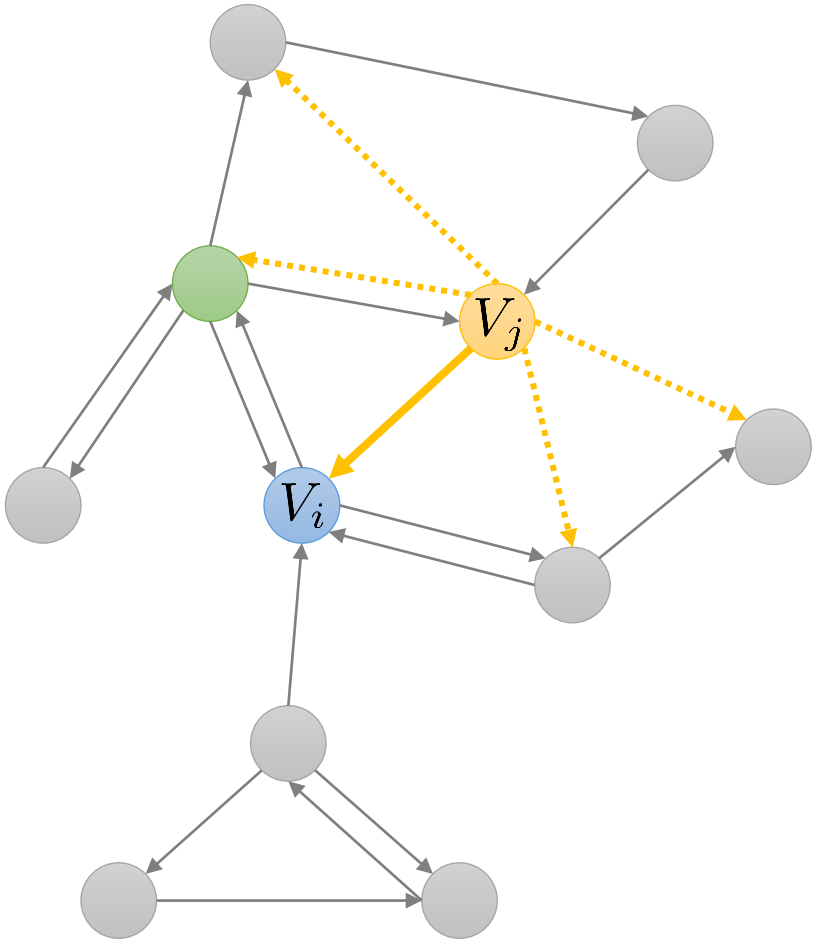


Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)

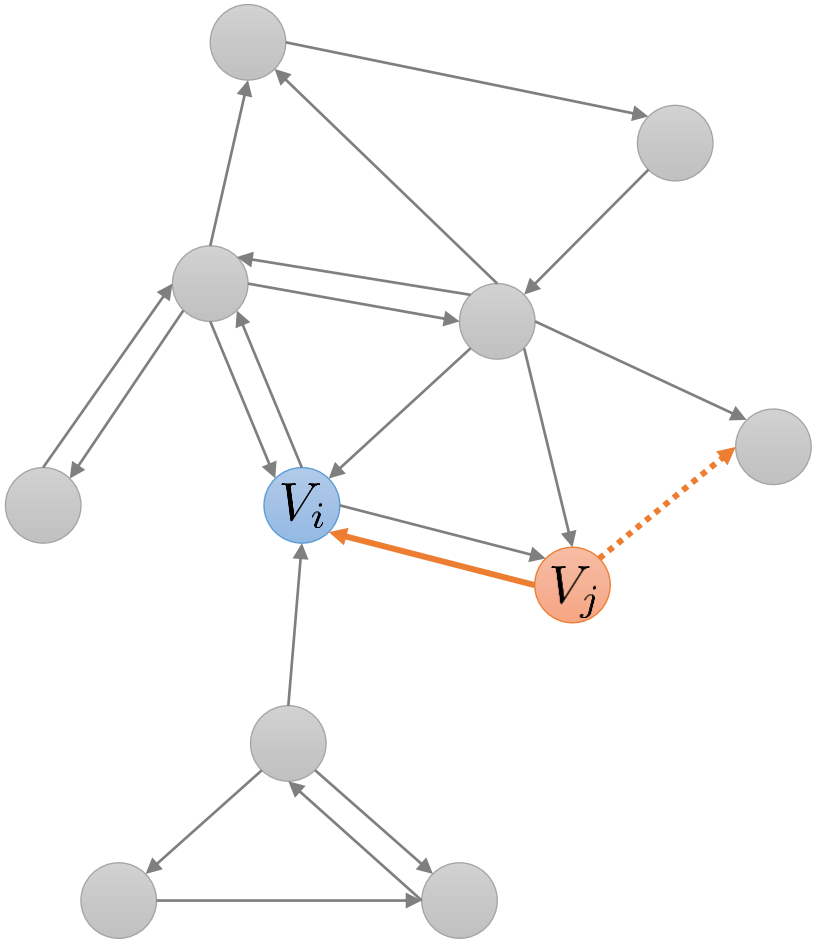


Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)

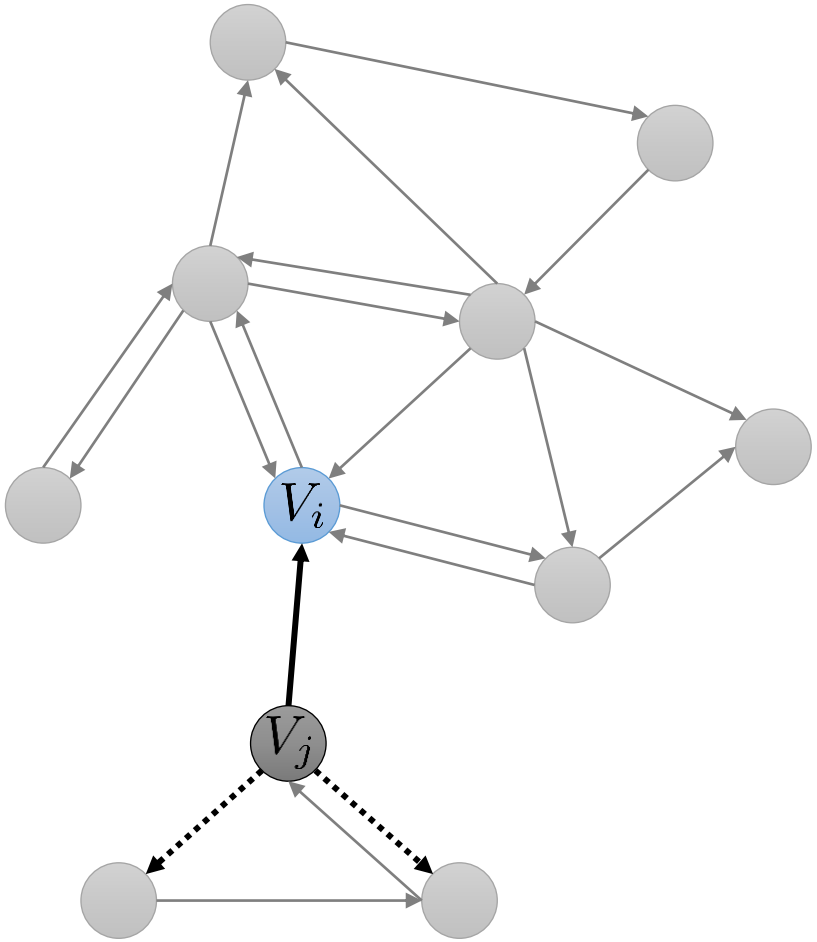


Graph-based Ranking

- PageRank
  - ✓ Let G = (V, E) be a directed graph with the set of vertices V and the set of edges E (a subset of V by V)
  - ✓ In(V_i): the set of vertices that point to it (predecessors)
  - ✓ Out(V_i): the set of vertices that vertex V*i* points to (successors)
  - ✓ The score of vertex V_i:

$$S(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

    - d is a damping factor allowing random work (usually set to 0.85)



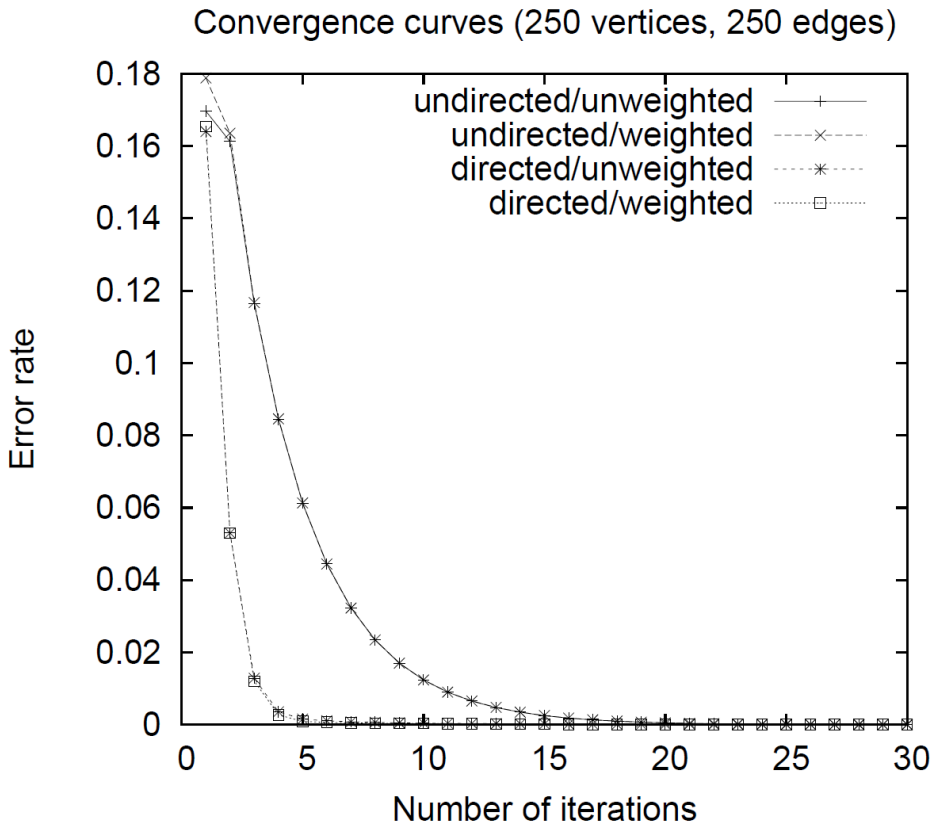
Graph-based Ranking

- PageRank
  - ✓ Starting from arbitrary values assigned to each node in the graph.
  - ✓ Computation iterates until convergence below a given threshold is achieved.
  - ✓ The final values obtained are not affected by the choice of the initial value, only the number of iterations to convergence may be different.

Graph-based Ranking

- PageRank(Page et al., 1999): Undirected graph and Weighted Graph
  - ✓ Undirected graph
    - The out-degree of a vertex is equal to the in-degree of the vertex
  - ✓ Weighted graph

$$WS(V_i) = \frac{1-d}{N} + d \times \sum_{j \in In(V_i)} \frac{\sum w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$



$$Error = S^{k+1}(V_i) - S^k(V_i)$$

(Mihalcea & Tarau, 2004)

Text as a Graph: TextRank

- Key issue: How to quantify the similarity between two sentences?

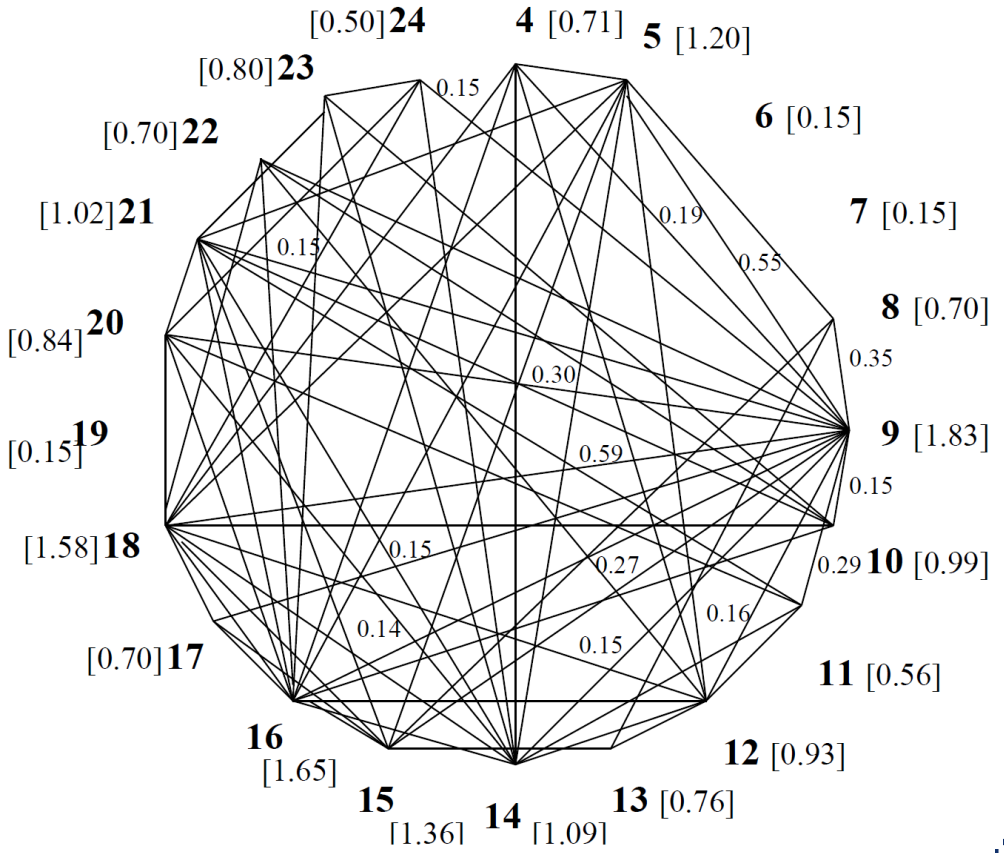
$$(\text{TextRank}) \quad Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \ \& \ w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

(Mihalcea & Tarau, 2004)

Text as a Graph: TextRank

- Graph representation of a text (**TextRank**)
  - ✓ Vertex: a lexical unit of text (sentence in this course)
  - ✓ Edge: similarity between the sentences

3: BC−HurricaineGilbert, 09−11 339
4: BC−Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.



(Mihalcea & Tarau, 2004)

Graph-based Text Summarization

- TextRank Summary Example

3: BC−HurricaineGilbert, 09−11 339
4: BC−Hurricaine Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14. Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricaine Saturday night.
15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

TextRank extractive summary
Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil De− fense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm. Strong winds associated with Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's south coast.

Manual abstract I
Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high wind and seas. Tropical storm Gilbert formed in the eastern Carribean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and in the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and down− graded to a tropical storm.

Manual abstract II
Tropical storm Gilbert in the eastern Carribean strenghtened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15 mph with a broad area of cloudiness and heavy weather with sustained winds of 75 mph gusting to 92 mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.

(Mihalcea & Tarau, 2004)

Text as a Graph: RexRank

- Key issue: How to quantify the similarity between two sentences?

$$\text{(TextRank)} \quad Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \ \& \ w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

(Mihalcea & Tarau, 2004)

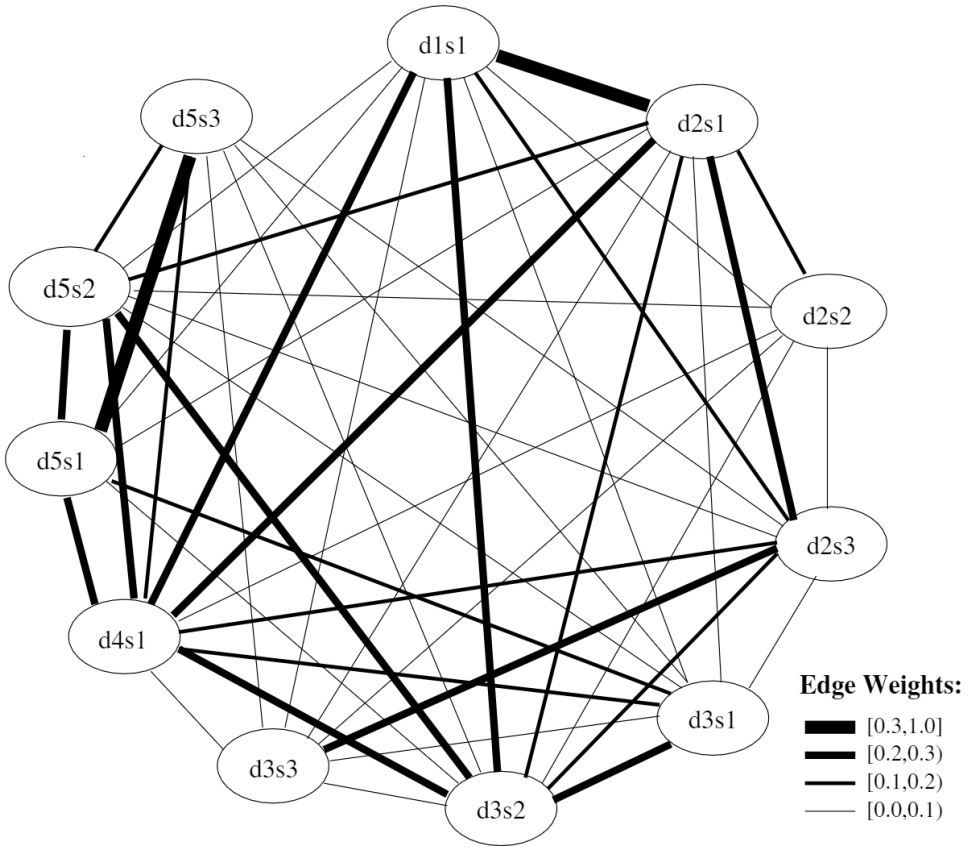
$$\text{(LexRank)} \quad Similarity(S_i, S_j) = \frac{\sum_{w \in S_i, S_j} \times tf_{w,S_i} \times tf_{w,S_j} \times (idf_w)^2}{\sqrt{\sum_{w \in S_i} (tf_{w,S_i} \times idf_w)^2} \times \sqrt{\sum_{w \in S_j} (tf_{w,S_j} \times idf_w)^2}}$$

(Erkan & Radev, 2004)

Text as a Graph: RexRank

- Graph representation of a text (**LexRank**)
  - ✓ Vertex: a lexical unit of text (sentence in this course)
  - ✓ Edge: similarity between the sentences

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.



(Erkan & Radev, 2004)

Text as a Graph

- Key issue: How to quantify the similarity between two sentences?

$$\text{(TextRank)} \quad Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \ \& \ w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

(Mihalcea & Tarau, 2004)

$$\text{(LexRank)} \quad Similarity(S_i, S_j) = \frac{\sum_{w \in S_i, S_j} \times tf_{w,S_i} \times tf_{w,S_j} \times (idf_w)^2}{\sqrt{\sum_{w \in S_i} (tf_{w,S_i} \times idf_w)^2} \times \sqrt{\sum_{w \in S_j} (tf_{w,S_j} \times idf_w)^2}}$$

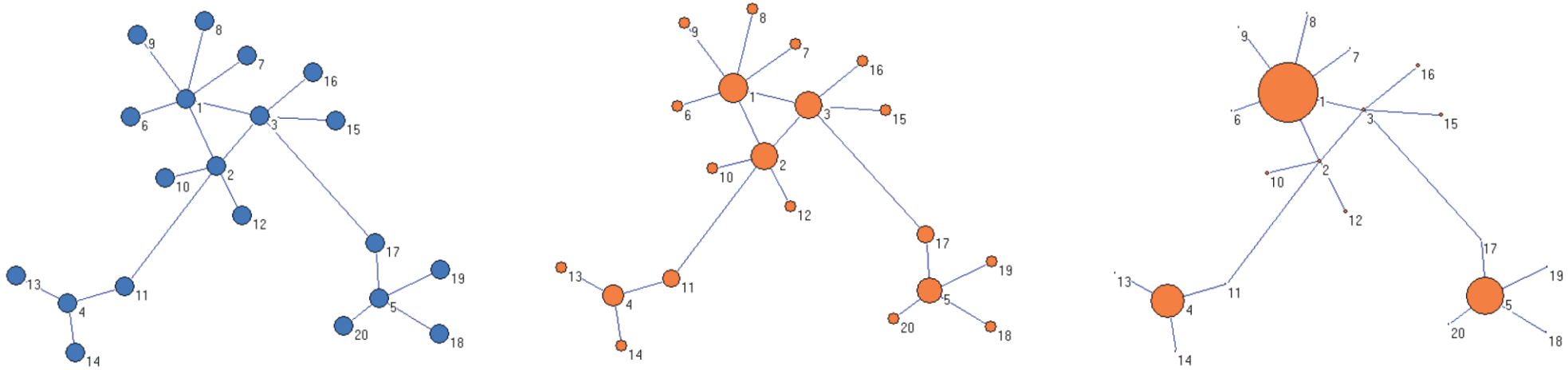
(Erkan & Radev, 2004)

- If we have sentence embedding vectors,

$$\text{(Embedding)} \quad Similarity(S_i, S_j) = \frac{S_i^T S_j}{|S_i| \times |S_i|}$$

DivRank

- Graph representation of a text (**DivRank**)
  - ✓ Vertex: a lexical unit of text (sentence in this course)
  - ✓ Edge: similarity between the sentences
  - ✓ Consider **Diversity**: the top ranked items are expected to contain as little redundant information as possible, cover as many aspects as possible, or be as independent as possible.



(a) An illustrative network. (b) Weighting with PageRank. (c) A diverse weighting.

(Mei et al., 2010)

DivRank

- Time-homogeneous random walk (PageRank)
  - ✓ Assume that the transition probabilities remain constant over time

$$p_T(v) = \sum_{(u,v)\in E} p(u,v)p_{T-1}(u)$$

$$p(u,v) = \begin{cases} \frac{1-d}{N} + d \times \frac{I[(u,v)\in E]}{deg(u)} & \text{if } deg(u) > 0 \\ \frac{1}{N} & \text{otherwise.} \end{cases}$$

DivRank

- In a real-world random walk process, it is reasonable to consider the change of transition probabilities over time

  ✓ **Vertex-reinforced random walks**

  - The transition probability to one state from others is reinforced by the number of previous visits to that state.

Distribution representing the prior preference of visiting vertex v

The number of times the walk has visited v up to time T

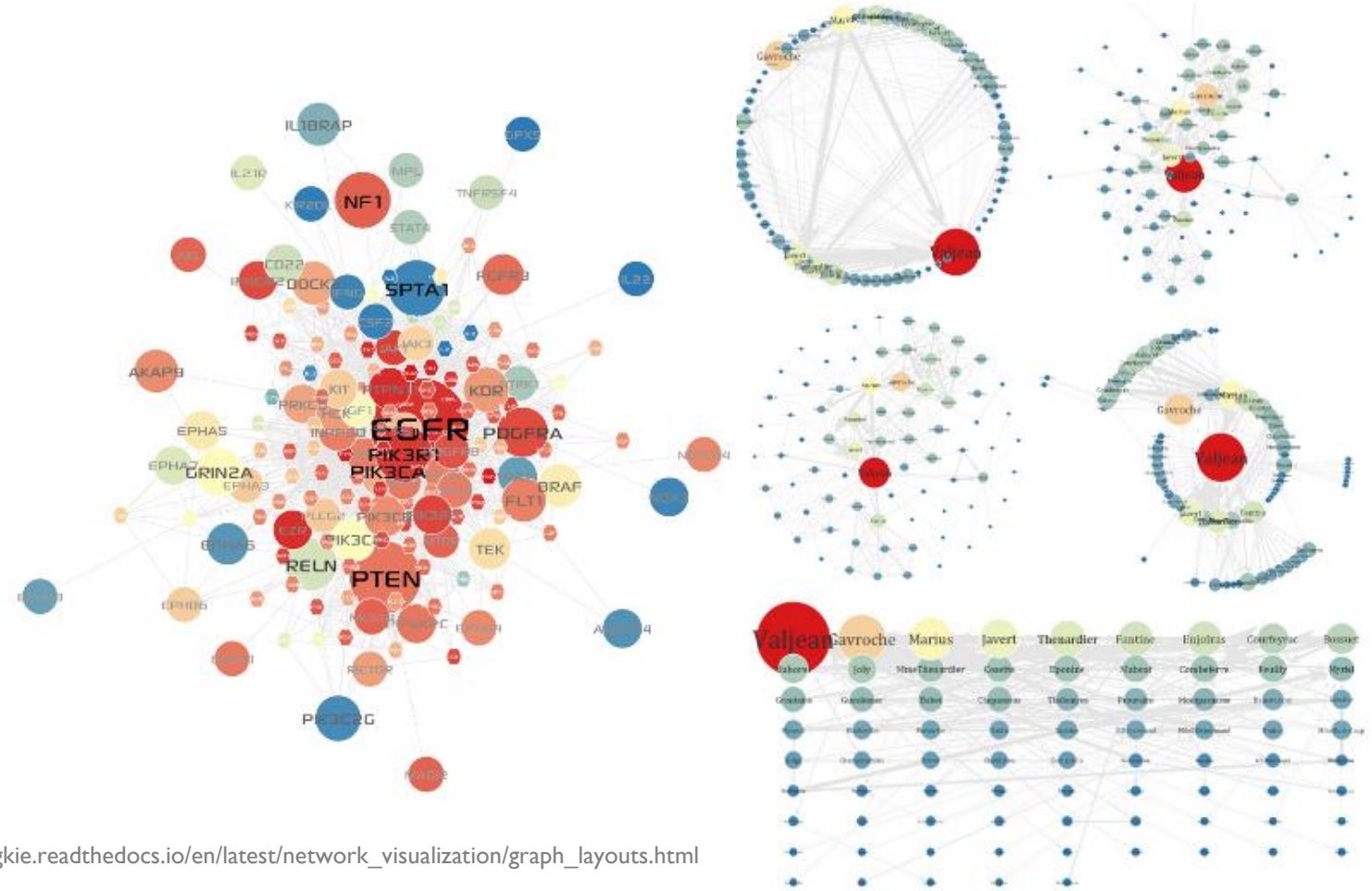
$$p_T(u,v) = (1-\lambda)\cdot p^*(v) + \lambda \cdot \frac{p_0(u,v)\cdot N_T(v)}{D_T(u)}$$

$$D_T(u) = \sum_{v \in V} p_0(u,v) N_T(v)$$

$$p_0(u,v) = \begin{cases} \alpha \cdot \frac{w(u,v)}{deg(u)} & \text{if } u \neq v \\ 1-\alpha & \text{if} u = v \end{cases}$$

(Optional) Graph

- Different layouts are possible for a single graph



https://mongkie.readthedocs.io/en/latest/network_visualization/graph_layouts.html

References

- Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
- Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, *22*, 457-479.
- Mei, Q., Guo, J., & Radev, D. (2010, July). Divrank: the interplay of prestige and diversity in information networks. *In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1009-1018).


ANY
questions?