Seokmin Kim's Statistic way.

김석민

<학부과정>

- 제 2회 L.Point big data competition - Be The LBA참가. (2015.11 ~ 2015.12)
- 동계 학부 연구생과정. (2016.12 ~ 2017.02)

<석사과정>

- R Package 구현 및 작성. (KNN Classification) (2019.03 ~ 2019.06)
- Auto-Encoding Variational Bayes. (2019.09 ~ 2019.10)
- Simulation of Markov Chain Monte Carlo. (2019. 10)
- 수온데이터 예측모형. (Prophet 모형 사용) (2019.11 ~ 2019.12)
- A Visualization Model Based on Adjacency data. 논문 구현. (2020. 04)
- Visualization of network data. (2020.03 ~ 2020.04)
- 드라마 '왕좌의 게임' 인물 관계 네트워크를 이용한 주요 인물 분석. (2020.05 ~ 2020.06)
- Analysis of Seoul air pollution data. (2020.08)
- Bayesian structural time series model and real data analysis(가제). (2020.01 ~)
- Evaluating mask policy of the Korean government using BSTS model. (2020.11 ~ 2020.12)
- 한국통계학회주관 제 28회 추계학술대회 포스터세션 2등.

- **Contents**

1. <드라마 '왕좌의 게임' 인물 관계 네트워크를 이용한 주요 인물 분석>

2. <Analysis of Seoul air pollution data>

<드라마 '왕좌의 게임' 인물 관계 네트워크를 이용한 주요 인물 분석>

1. 데이터 설명

- 미국 드라마 "왕좌의 게임" 등장인물 간 상호작용을 그래프 형태로 나타낸 데이터.
- 비 방향성 및 가중치 그래프로 제공되나, Link community 알고리즘을 위해 비 가중치를 가정하고 분석.

2. 분석 목적.

- 문장 단위로 연결이 정의되는 소설보다 현실 네트워크와 유사한 드라마 네트워크 데이터를 바탕으로 인물 간의 네트워크를 분석.
- Centrality measure를 이용한 주요 인물 탐색과 Link community를 통해 주요 인물 및 커뮤니티를 탐색

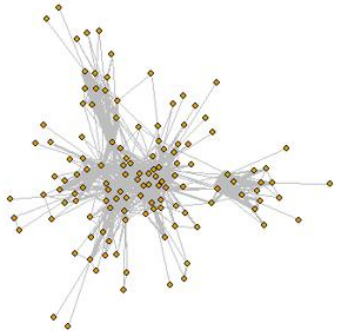
3. 분석 방법:Link community (Ahn et al. 2010)

- 노드가 아닌 엣지의 커뮤니티를 탐색하는 알고리즘.
- 커뮤니티의 계층적 구조를 파악할 수 있음.
- 커뮤니티간 겹침(overlap)이 발생할 수 있기 때문에, 하나의 노드가 여러 커뮤니티에 포함 가능.

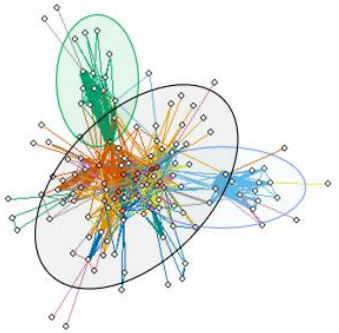
4. 데이터 분석

- Season 1 링크커뮤니티 시각화
  - 커뮤니티1(검은색): 시즌1 주요 줄거리에 속한 인물들로 구성된 크고 작은 커뮤니티들.
  - 커뮤니티2(초록색): '존'이 속하는 '나이트 워치'라는 집단에 포함된 커뮤니티.
  - 커뮤니티3(하늘색): '도트락'부족과 '타르가리엔'가문 인물들로 구성된 커뮤니티.

The graph of Season 1



The graph of Season 1 with link communities



5. 결론

- 주요 인물로 예상했던 인물들이 Centrality measure 분포의 상위에 위치함을 확인함.
- Centrality measure로 뽑힌 중요 인물과 Overlap count로 뽑힌 중요 인물이 매우 유사한 것으로 나옴.
- Overlap count도 주요 인물 탐색에 활용될 수 있을 것으로 생각됨.
- 다수의 작은 커뮤니티에 겹쳐져 있는 인물이 높은 Overlap count 값을 가져 Centrality 기준에서 높은 위치에 포함되지 않았던 인물이 종종 등장하는 것을 확인
- 실제 사회 네트워크에서는 연결의 유무 뿐만 아니라 연결의 강도가 매우 중요한 정보로 활용되기 때문에, 가중치를 갖는 그래프를 고려할 수 있도록 확장한 Link similarity를 사용하면 이러한 문제를 해결할 수 있을 것으로 기대.

<Analysis of Seoul air pollution data>

1. 데이터 설명

- 2017년 01,01부터 2019년 12,31까지 서울특별시 보건환경연구원 대기오염측정망시스템에서 제공하는 서울특별시 25개구에 대한 대기오염 시간별 측정정보.
- Measurement info, Measurement item info, Measurement station, Measurement summary라는 4개의 데이터 셋으로 이루어져 있음.
- Meausrement info 데이터는 3885066개의 obs와 5개의 변수가 존재.

2. 데이터 전처리

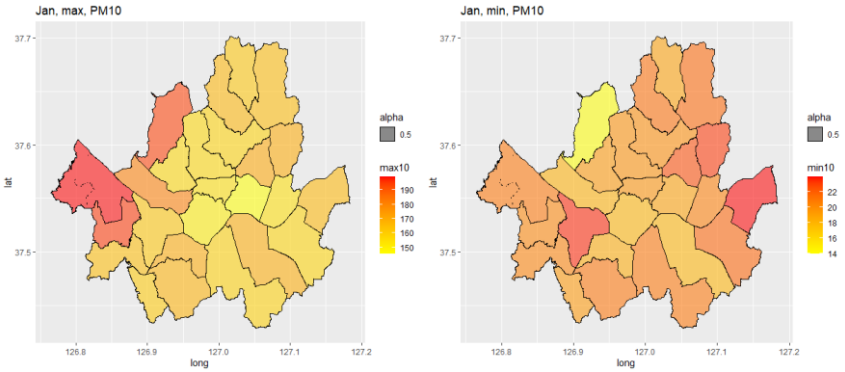
- 1차 데이터 변환결과 다음과 같은 missing value가 존재.

SO2	NO2	CO	O3	PM10	PM2.5	Total
26256	24436	27216	25880	32592	29842	166228

- 측정 기기오류로 잘못 측정된 데이터 missing value로 재처리.
- Missing value 값을 채우기 위해 R의 ImputeTS 패키지의 na_kalman함수 사용.

3. 분석 및 시각화

- 특정월의 일평균 최대, 최소값을 서울지도에 시각화 표현.
-)PM 10에 대한 일평균 최대, 최소값 시각화



5. 분석방법

- Training set[2017.01.01 ~ 2018.09.30], Test set[2018.10.01 ~ 2018.12.31] Predict set[2019.01.01 ~ 2019.01.03]으로 설정.
- Bayesian structural time series모형을 사용하기위해 R의 "bsts"패키지를 사용.
- 분석결과 spike and slab prior를 통해 미세먼지에 영향을 끼치는 변수들을 파악.

6. 결론 및 향후과제

- Bayesian structural time series모형은 nowcasting의 모형으로 적합하나 단기예측에도 상당히 좋은 결과를 얻을 수 있다는 것을 알게 됨.
- 상태공간모형의 latent variable에 대해 조금 더 연구가 필요함,
- 장기예측에도 사용이 가능한지 연구가 필요함.