

위치 기반 소셜 네트워크에서 공간 근접 커뮤니티 탐지를 위한 엔트로피 가중 적응형 라벨 전파

오도열* 김혜원* 김다희 김정훈†
울산과학기술원

{ohdoyoel, hyewon.kim, dahee, junghoon.kim}@unist.ac.kr

Entropy-Guided Adaptive Label Propagation for Location-Aware Graph Clustering

Doyeol Oh* Hyewon Kim* Dahee Kim Junghoon Kim†
Ulsan National Institute of Science & Technology

요약

최근 IT 기술의 발전과 함께 소셜 네트워크에서 활용 가능한 메타데이터의 범위가 크게 확장되고 있다. 특히 위치 기반 소셜 네트워크(Location-Based Social Network, LBSN)는 사용자 간의 연결 구조뿐 아니라 위치 정보를 포함하여 사회적 관계를 더욱 풍부하게 설명할 수 있어 많은 관심을 받고 있다. 본 연구에서는 이러한 맥락에서 네트워크의 구조적 정보와 위치적 정보를 동시에 고려할 수 있는 커뮤니티 탐지 방법을 제안한다. 구체적으로, 확장성이 높다고 알려진 라벨 전파(Label Propagation, LP) 알고리즘을 기반으로, 이웃 노드의 라벨 분산 정도를 정량화하는 엔트로피를 활용하여 구조적 신뢰도를 계산하고 이에 따라 구조·위치 정보의 비중을 적응적으로 조정하는 방식을 설계하였다. 제안된 방법은 실제 LBSN 데이터셋을 대상으로 한 실험을 통해 기존 기법 대비 안정적이고 일관된 커뮤니티 탐지 성능을 달성함을 확인하였다.

1. 서론

최근 IT 기술의 발전과 함께 소셜 네트워크 서비스는 단순한 연결 구조를 넘어 다양한 메타데이터를 포함하는 복합 시스템으로 발전하고 있다. 특히 위치 기반 소셜 네트워크(Location-Based Social Network, LBSN) [1]는 GPS 좌표와 같은 공간 정보를 통해 사회적 관계의 맥락을 풍부하게 표현할 수 있어 많은 주목을 받고 있다. 위치 정보는 물리적 제약을 수반하며, 이는 사회적 상호작용의 가능성에 직접적인 영향을 미친다. 예를 들어 동일 지역 내 사용자들은 교통, 날씨, 지역적 사건을 공유하면서 관계가 강화되는 지역성(locality preference)을 보이는 것으로 알려져 있다 [2]. 따라서 LBSN은 기존의 구조적 연결만으로는 파악하기 어려운 사회적 관계를 분석하는 데 중요한 기반을 제공한다.

소셜 네트워크 분석에서 커뮤니티 탐지는 오랫동안 핵심 과제로 다루어져 왔다 [3]. 대표적인 알고리즘인 라벨 전파(Label Propagation, LP) [4]는 다수결 원리를 기반으로 효율적으로 커뮤니티를 탐지할 수 있지만, 브릿지 노드가 존재할 경우 뚜렷한 구조가 하나로 합쳐지는 문제 [5]가 발생한다(그림 1a). 이를 개선하기 위해 Jaccard 유사도를 활용한 방법 [6]이 제안되었으나(그림 1b), 여전히 구조적 연결만을 강조하며 위치 정보는 배제된 상태이다. 반대로 단순히 공간적 근접성만을 반영할 경우, 구조적으로 밀집된 관계를 무시하고 거리에 따라 커뮤니티를 불필요하게 분리하는 문제가 나타난다(그림 1c). 이러한 한계는 구조적 정보와 위치 정보를 함께 고려하는 접근이 필요함을 시사한다.

그러나 구조적 정보가 항상 동일한 신뢰도를 갖는 것은 아니다. 업무 중심 지역에서는 조직 내·외부의 잦은 상호작용으로 인해 구조적 신호가 뚜렷하게 나타나는 반면, 관광이나 여가 중심 지역에서는 관계가 일시적이고 유동적이어서 구조적 신호가

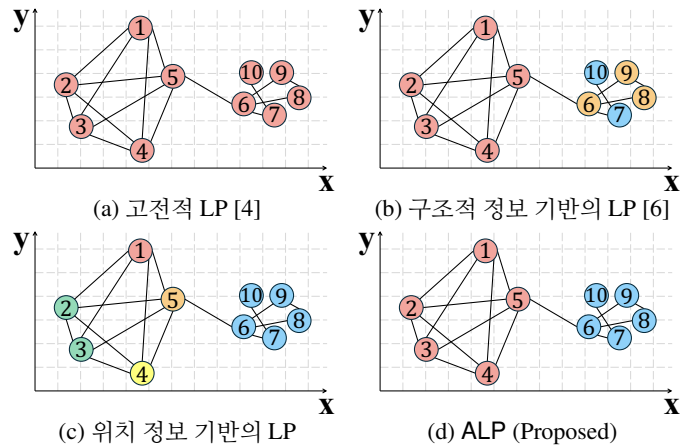


그림 1: 다양한 LP 방식에 대한 예제

상대적으로 약하다는 특징이 있다. 이 경우 구조적 정보보다 위치 정보가 중요한 단서로 작용할 수 있다. 따라서 커뮤니티 탐지를 위해서는 상황에 따라 구조와 위치의 기여도를 동적으로 조절할 수 있는 접근이 필요하다.

본 논문에서는 이러한 문제를 해결하기 위해, 구조적 신뢰도를 정량화하고 이에 따라 구조·위치 정보의 비중을 적응적으로 조정하는 위치 기반 적응형 라벨 전파(Adaptive Label Propagation, ALP)를 제안한다. 제안된 접근은 구조적 신뢰도가 높은 영역에서는 구조 기반 커뮤니티를, 구조적 신호가 불안정한 영역에서는 위치 정보를 활용한 커뮤니티를 자연스럽게 식별한다(그림 1d). 이를 통해 지역적 특성과 사회적 맥락을 균형있게 반영한 보다 안정적이고 정확한 커뮤니티 탐지가 가능하다.

2. 이론적 배경

이 장에서는 ALP의 기본 개념을 설명한다. LBSN $G = (V, E, C)$ 에서 V 는 노드 집합, E 는 엣지 집합, C 는 각 노드별 위치 정보를 나타낸다. 구조적·위치적 정보 기반 라벨 전파를 위

*These authors contributed equally to this work.

†Corresponding author.

한 유사도 지표들은 아래와 같이 정의된다.

구조적 유사도는 두 노드의 구조적 유사성을 측정하는 지표로 그래프 분석에서 널리 사용되는 Jaccard 유사도를 활용한다.

정의 1 (구조적 유사도). 노드 쌍 $i, j \in V$ 가 주어졌을 때, 구조적 유사도는 다음과 같이 정의된다: $sim_{str}(i, j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$. 여기서 $N(i)$ 는 i 의 이웃 집합이다.

유사도가 0이 되는 경우를 방지하기 위해 자기 자신을 포함하여 계산한다. 이때 두 노드가 공유하는 이웃이 많을 수록 높은 구조적 유사도를 갖도록 한다.

위치적 유사도는 노드 간 지리적 근접성을 측정하기 위한 지표로, 두 노드의 위치가 가까울수록 높은 값을 갖도록 한다.

정의 2 (위치적 유사도). 두 노드 $i, j \in V$ 가 주어졌을 때, 위치적 유사도는 다음과 같이 정의된다: $sim_{loc}(i, j) = 1 - \frac{dist(c_i, c_j)}{\max_{u, v \in V} (dist(c_u, c_v))}$. 이때 c_i 는 i 의 위치를, $dist(c_i, c_j)$ 는 i 와 j 의 위치 간 하버사인 거리(Haversine distance)를 나타낸다.

위에서 정의한 구조적·위치적 유사도를 활용하여 라벨 전파 과정에서 구조 및 위치 정보를 통합적으로 반영한다. 이때 이웃 라벨의 일관성 정도에 따라 이웃 정보의 신뢰도가 달라진다. 대부분의 이웃이 동일한 라벨을 가진다면, 해당 노드의 커뮤니티 소속에 대한 신호가 명확하므로 구조적 정보의 신뢰도가 높다고 간주할 수 있다. 반대로 이웃 라벨이 여러 커뮤니티로 고르게 분산되어 있다면, 적절한 라벨 선택에 대한 불확실성이 커지게 된다. 이러한 이웃 정보의 신뢰도는 라벨 엔트로피를 통해 정량화된다.

정의 3 (라벨 엔트로피). 노드 $i \in V$ 가 주어졌을 때, 라벨 엔트로피는 다음과 같이 정의된다: $H_i = -\sum_{l \in \mathcal{L}} p_i(l) \log p_i(l)$. 여기서 $p_i(l)$ 은 i 의 이웃 중 라벨이 l 인 노드의 비율이다.

라벨 엔트로피는 노드 주변의 라벨 분포 일관성을 측정하는 지표로, 엔트로피 값이 낮을수록 대다수 이웃이 동일한 라벨을 갖는다. 이는 해당 노드의 커뮤니티 소속에 대한 구조적 신호가 명확함을 의미하므로, 이를 통해 구조적 정보에 대한 신뢰도를 평가할 수 있다. 이를 활용하여 구조적·위치적 정보의 비중을 적응적으로 조절하는 적응 유사도를 아래와 같이 정의한다.

정의 4 (적응 유사도). 노드 $i \in V$ 와 이웃 노드 $j \in N(i)$ 가 주어졌을 때, 적응 유사도는 다음과 같이 정의된다: $sim(i, j) = \alpha_j \cdot sim_{str}(i, j) + (1 - \alpha_j) \cdot sim_{loc}(i, j)$. 여기서 적응 가중치는 $\alpha_j = 1 - \frac{H_j}{\log |\mathcal{L}_j|}$ 로 정의된다.

α 가 클수록 라벨의 분포가 일관적이어서 구조적 신호의 신뢰도가 높아 구조적 정보의 비중이 증가하고, 작을수록 위치적 정보에 더 의존하도록 설계되었다. 특히 $\alpha = 1.0$ 일 때 구조적 유사도만을 고려하기에 Wang et al. [6]의 연구 방법과 동일하다.

3. 문제 정의

본 장에서는 2장에서 제시한 개념들을 통해 기존의 LP를 확장한 위치 기반 적응형 라벨 전파 알고리즘 (ALP)을 제안한다.

Algorithm 1: Adaptive Label Propagation (ALP)

Input: LBSN $G = (V, E, C)$ and the maximum number of iterations T

Output: Detected community labels $\mathcal{L} = \{\mathcal{L}_i \mid i \in V\}$

- 1 $\alpha \leftarrow \emptyset$; $sim \leftarrow \emptyset$; $iterations \leftarrow 0$;
- 2 $\mathcal{L} \leftarrow \text{InitLabels}(V)$;
- 3 $sim_{str} \leftarrow \text{ComputeSim}_{str}(G)$;
- 4 $sim_{loc} \leftarrow \text{ComputeSim}_{loc}(G)$;
- 5 **while** labels have not converged **and** $iterations < T$ **do**
- 6 $iterations \leftarrow iterations + 1$;
- 7 **foreach** $i \in V$ **do**
- 8 $\alpha_i \leftarrow \text{UpdateWeights}(G, \mathcal{L}_i)$;
- 9 **foreach** $(i, j) \in E$ **do**
- 10 $sim(i, j) \leftarrow \text{ComputeSim}(\alpha_j, sim_{str}(i, j), sim_{loc}(i, j))$;
- 11 **foreach** $i \in V$ **do**
- 12 $\mathcal{L}_i \leftarrow \text{UpdateLabels}(i, sim, \mathcal{L})$;
- 13 **return** \mathcal{L}

문제 정의 1 (위치 기반 적응형 라벨 전파 알고리즘 (Adaptive Label Propagation, ALP)). LBSN $G = (V, E, C)$ 가 주어졌을 때, 각 노드 $i \in V$ 는 자신과 이웃 노드 $j \in N(i)$ 간의 적응 유사도 $sim(i, j)$ 를 기반으로 라벨을 반복적으로 갱신하여 최적의 커뮤니티 구조를 찾는다. 라벨 갱신 규칙은 다음과 같다:

$$\mathcal{L}_i^{(t+1)} = \arg \max_{l \in \mathcal{L}} \sum_{j \in N(i)} sim(i, j) \cdot \mathbb{I}[\mathcal{L}_j^{(t)} = l]$$

여기서 $\mathbb{I}[\mathcal{L}_j^{(t)} = l]$ 은 지시 함수로, 가능한 라벨 집합 \mathcal{L} 에 대해, 이웃 j 의 라벨이 l 일 경우 1, 그렇지 않으면 0의 값을 갖는다.

ALP는 각 라벨 후보 l 에 대해 이웃 노드들 중 현재 라벨이 l 인 노드들의 적응 유사도 합을 계산하고, 그 합이 가장 큰 라벨로 갱신한다. 이를 통해 기존의 단순 다수결 기반 라벨 전파를 확장하여 위치 정보를 통합함으로써, 구조적으로 밀접하면서도 공간적으로 응집력 있는 커뮤니티를 효과적으로 식별할 수 있다.

4. 알고리즘

본 장에서는 기존의 LP를 기반으로 구조적·위치적 유사도를 적응적으로 조합하여 라벨을 전파하는 ALP의 계산 과정을 제시한다. 이 알고리즘은 각 노드에 대해 이웃 노드의 라벨 엔트로피가 낮을 때는 구조적 유사도에 따른 전파를 우선시하고, 그렇지 않을 경우에는 위치적 유사도에 더 의존하도록 동작한다.

알고리즘 설명. 알고리즘은 두 단계로 구성된다: (1) 초기화: LBSN G 의 각 노드에 라벨을 부여하고, 모든 엣지에 대해 구조적·위치적 유사도를 계산한다. (2) 라벨 전파: 각 반복 단계에서 노드의 라벨 엔트로피로부터 산출한 구조적 신뢰도를 이용하여 두 유사도를 결합한 적응 유사도를 생성한다. 각 노드는 이를 바탕으로 가중 다수결 방식으로 라벨을 갱신하며, 이 과정은 라벨이 완전히 수렴할 때까지 반복된다.

시간 복잡도 분석. 초기화 단계에서는 라벨 초기화에 $O(|V|)$, 구

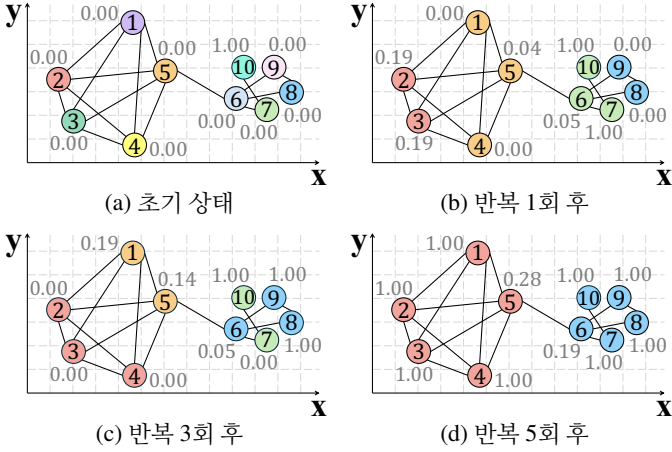


그림 2: ALP 계산 과정. 회색 수치는 각 노드의 α 값을 의미한다.

조적 유사도 계산에 $O(d_{max}|E|)$, 그리고 위치적 유사도 계산에 $O(|E|)$ 의 시간이 소요되며, 여기서 d_{max} 는 그래프 최대 이웃 수를 나타낸다. 라벨 전파 단계에서 내부의 각 반복이 $O(|E|)$ 의 비용을 가지고, 전체 반복 횟수 T 에 따라 총 비용은 $O(T|E|)$ 이다. 따라서 전체 시간 복잡도는 $O(\max(d_{max}, T) \cdot |E|)$ 이다.

예제 1. 그림 2는 LBSN에서 ALP 알고리즘이 수행되는 과정을 보여주며, 각 노드 옆 회색 수치는 적응 가중치 α 를 나타낸다. 초기에는 대부분의 노드가 서로 다른 라벨을 가져 라벨 엔트로피가 높고, 이에 따라 적응 가중치가 작아 주로 위치적 유사도에 의존해 인접 노드의 라벨을 따른다(그림 2a). 라벨 전파가 진행되면서 인접 노드의 라벨이 변하면 각 노드의 적응 가중치도 함께 갱신된다(그림 2b). 이후, $\alpha = 0.19$ 를 갖는 노드 2와 3의 영향으로 위치는 멀지만 구조적으로 밀집된 노드 4의 라벨이 바뀌고, $\alpha = 0$ 으로 위치적 유사도만을 고려하는 노드 8과 9의 영향으로 위치가 가까운 노드 6의 라벨이 갱신된다(그림 2c). 라벨이 더 이상 변하지 않을 때까지 반복하면, 최종적으로 {1, 2, 3, 4, 5}와 {6, 7, 8, 9, 10}의 두 커뮤니티가 형성된다(그림 2d).

5. 실험

제안한 방법론의 유효성을 평가하기 위해 다양한 실험을 수행하였으며, 이를 통해 다음과 같은 평가 질문(Experimental Questions, EQ)에 대한 답을 제시하고자 한다.

- **EQ1. 모델 성능 평가:** 다양한 평가 지표를 통해 제안한 모델의 클러스터링 품질과 공간적 일관성 간 균형을 입증한다.
- **EQ2. 수렴 효율성 평가:** ALP의 수렴 속도를 기존 기법과 비교하여, 반복 횟수 측면에서 효율성을 검증한다.
- **EQ3. 적응 동작 분석:** 반복에 따른 적응 가중치 α 의 분석을 통해 구조-위치간의 동적 조절 능력을 평가한다.
- **EQ4. 사례 연구:** 실제 데이터에서 LP와 ALP의 커뮤니티 탐지 결과를 시각화하여 비교 분석한다.

데이터셋. 제안한 모델을 평가하기 위해 사용자 체크인 정보와 친구 관계를 포함하는 LBSN 데이터를 사용하였다. 위치 정보가

표 1: Real-world datasets

Dataset	$ V $	$ E $	Avg. Deg.
Brightkite [7]	50,686	388,180	15.32
Gowalla [7]	107,069	913,536	17.06

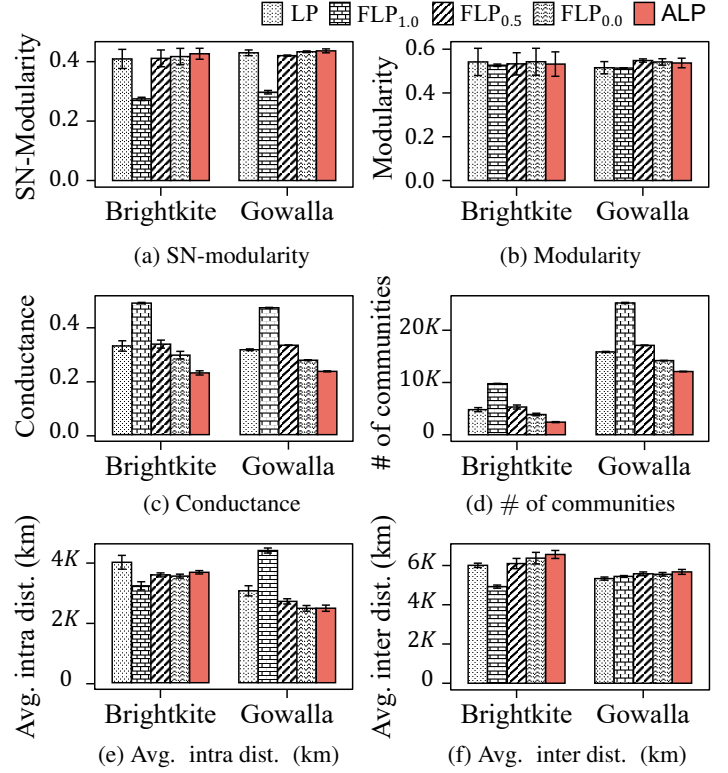


그림 3: EQ1. 모델 성능 평가

존재하지 않는 노드는 전처리 과정을 통해 제외되었으며, 각 사용자의 위치 정보는 가장 최근의 체크인 정보를 활용하였다. 자세한 데이터 정보는 표 1에서 확인할 수 있다.

비교 모델. ALP의 성능은 고전적인 라벨 전파 방법(LP) [4]과, α 값을 고정한 ALP의 변형인 고정 라벨 전파 (Fixed Label Propagation, FLP)와 비교하였다. FLP은 구조적 유사도만 반영하는 FLP_{1.0} [6], 구조적-위치적 유사도를 동일하게 반영하는 FLP_{0.5}, 위치적 유사도만 반영하는 FLP_{0.0}을 포함한다.

실험 설정. ALP의 랜덤 초기화와 순서 의존성을 고려하여 모든 실험 결과는 50회 반복 수행한 값의 평균과 표준편차로 제시한다.

EQ1. 모델 성능 평가. 그림 3은 실제 데이터셋에서 ALP와 기존 방법들의 성능을 여섯 가지 지표로 비교한 결과이다. ALP는 위치 기반 modularity 지표인 SN-modularity [8]에서 상대적으로 높은 값을, conductance에서 낮은 값을 보여 우수한 성능을 나타냈다(그림 3a, 3c). 여기서 낮은 conductance 값은 커뮤니티 간 경계가 뚜렷함을 의미한다. Modularity [9]의 경우, ALP가 위치 정보를 함께 고려하는 특성으로 인해 일부 구조적 관계가 희석되어 다소 낮은 값을 보였으나, 다른 모델과 유사한 수준을 유지하였다(그림 3b). 또한 ALP는 커뮤니티 개수가 적어 과분할을 억제하였고(그림 3d), 커뮤니티 내부 평균 거리는 비슷하거나 더 낮은 수준을 유지하였다(그림 3e). 반대로 커뮤니티 간 평균 거리는 더 크게 나타나(그림 3f), ALP가 구조적-위치적 정보를 균형 있게 반

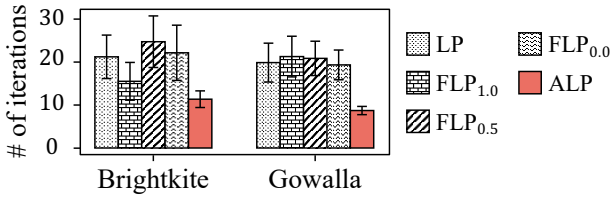


그림 4: EQ2. 수렴 효율성 평가

영하여 우수한 커뮤니티 탐지를 수행했음을 보여준다.

EQ2. 수렴 효율성 평가. 그림 4는 비교 모델들과 ALP가 각각 수렴하기까지 필요한 반복 횟수를 비교한 결과이다. ALP는 다른 방법들에 비해 반복 횟수가 현저히 적어, 앞서 4장에서 분석한 시간 복잡도에서 반복 수 T 가 실제로 크지 않음을 확인할 수 있다. 또한 기존 LP 계열 모델들은 무작위성에 따라 매 실행마다 수렴 속도의 편차가 크게 나타나는 반면, ALP는 변동 폭이 상대적으로 작아 결과의 안정성이 높다. 이는 ALP가 노드 업데이트 순서에 크게 의존하지 않고, 구조적·위치적 정보를 동시에 활용함으로써 보다 강건한 수렴 특성을 갖는다는 점을 시사한다.

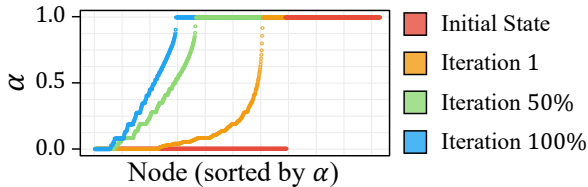


그림 5: EQ3-1. 적응 동작 분석 (Brightkite 데이터셋)

EQ3-1. 적응 동작 분석 (Brightkite 데이터셋). 그림 5은 Brightkite 데이터셋에서 반복적인 라벨 전파 과정에 따른 α 값의 분포 변화를 보여준다. 초기 단계에서는 모든 노드가 서로 다른 라벨을 가지고 있어 구조적 신뢰도가 거의 없으므로, 이웃이 하나뿐인 노드만 $\alpha = 1$ 을 갖고, 나머지는 $\alpha = 0$ 에 머무른다. 이 시기에는 위치 정보가 주요 의사결정의 기준이 되며, 반복이 진행됨에 따라 구조적 신뢰도가 점차 높아져 노드들의 α 값이 점차 증가하는 경향을 보인다.

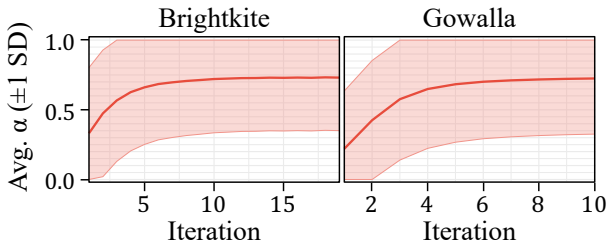
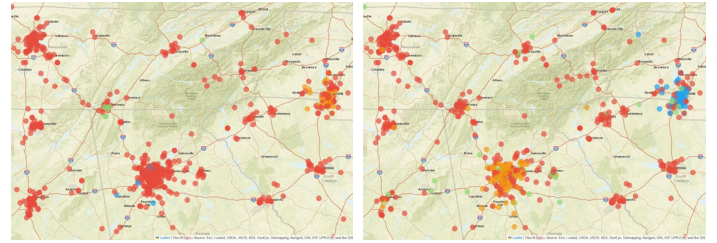


그림 6: EQ3-2. 적응 동작 분석

EQ3-2. 적응 동작 분석. 그림 6은 반복에 따른 α 의 평균 및 표준편차 변화를 나타낸다. 전반적으로 반복이 진행될수록 α 값의 평균은 증가하고 표준편차는 감소하는 경향을 보인다. 이는 ALP가 라벨 전파를 거듭할수록 이웃 간 라벨의 불확실성을 줄이고, 구조적 신뢰도를 점진적으로 높여 안정화되는 과정을 반영한다.

EQ4. 사례 연구. 그림 7는 Brightkite 데이터셋에서 LP와 ALP를 적용한 결과를 보여준다. 동일한 커뮤니티는 같은 색으로 표시하였으며, 미국 동남부의 내슈빌, 아틀란타, 샬럿 지역을 대상으로



(a) LP의 시각화 결과

(b) ALP의 시각화 결과

그림 7: EQ4. 사례 연구 (Brightkite 데이터셋)

시각화하였다. 그림 7a의 LP 결과는 구조적 연결만을 반영하기 때문에 실제 지리적 분포와 무관하게 분산된 도시들이 동일한 클러스터로 묶이는 경향이 있음을 보여준다. 반면 그림 7b의 ALP 결과는 내슈빌, 아틀란타, 샬럿 지역이 각각 분리된 클러스터로 나타나며, 위치 정보를 함께 고려함으로써 구조적·위치적으로 적합한 클러스터를 형성함을 보여준다.

6. 결론

본 연구는 LBSN에서 라벨 전파 방법을 확장하기 위해 엔트로피 기반 적응 가중치를 활용한 적응형 라벨 전파 알고리즘을 제안하였다. ALP는 각 노드의 구조적 신뢰도를 정량화하여 구조적 유사도와 위치적 유사도의 비중에 동적으로 조절함으로써, 안정적인 커뮤니티 탐지를 가능하게 한다. 또한 제안한 알고리즘은 실제 데이터 기반의 실험을 통해 기존 커뮤니티 탐지 기법 대비 우수한 성능을 입증하였다.

Artifacts. <https://github.com/ohdoyoel/alp.git>

감사의 글. 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단(IRIS RS-2025-00523578) 및 2025년도 UNIST(울산과학기술원) 학부생 융합연구프로젝트(UIRP)의 지원으로 수행되었음.

참고 문헌

- [1] Y. Zheng, "Location-based social networks: Users," in *Computing with spatial trajectories*, pp. 243–276, Springer, 2011.
- [2] J. Miller, "Activity-based intelligence: Principles and applications," *Defense and Intelligence Journal*, vol. 22, no. 1, pp. 1–12, 2013.
- [3] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.
- [4] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, p. 036106, 2007.
- [5] J. Hou Chin and K. Ratnavelu, "A semi-synchronous label propagation algorithm with constraints for community detection in complex networks," *Scientific Reports*, vol. 7, no. 1, p. 45836, 2017.
- [6] M. Wang, X. Cai, Y. Zeng, and X. Liang, "A community detection algorithm based on jaccard similarity label propagation," in *IDEAL*, pp. 45–52, 2017.
- [7] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *SIGKDD*, pp. 1082–1090, 2011.
- [8] J. Hannigan, G. Hernandez, R. M. Medina, P. Roos, and P. Shakarian, "Mining for spatially-near communities in geo-located social networks," in *AAAI Fall Symposia*, 2013.
- [9] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *PRE*, vol. 69, no. 2, p. 026113, 2004.