
Fusion de données géographiques par alignement de graphes: application aux réseaux d'assainissement

Omar Et-targuy¹ Carole Delenne¹ Salem Benferhat²

¹ IUSTI, Aix Marseille Université – CNRS, Marseille

² CRIL, CNRS UMR 8188, Université d'Artois

Premier Congrès Eau & IA — 2025



Images



Cartes



Vidéos

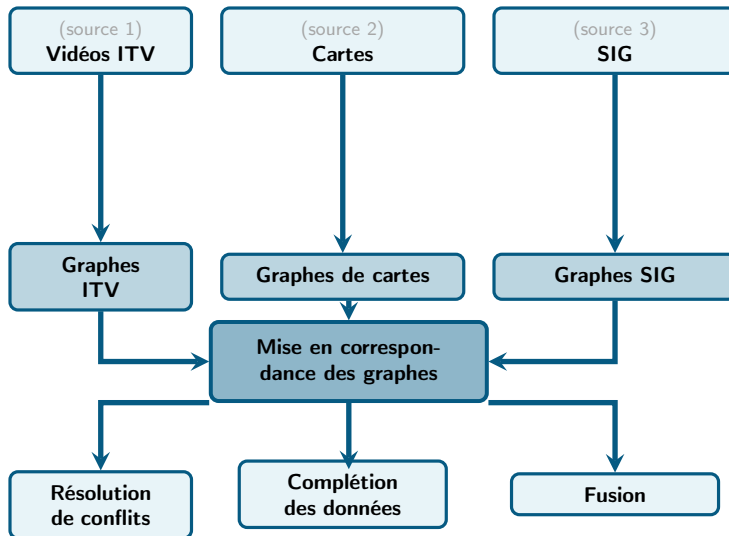
0.16	0.000000	
	0.000000	AC
	0.000000	PVC
	0.000000	PVC
	0.000000	AC
	0.000000	

Bases de données tabulaire

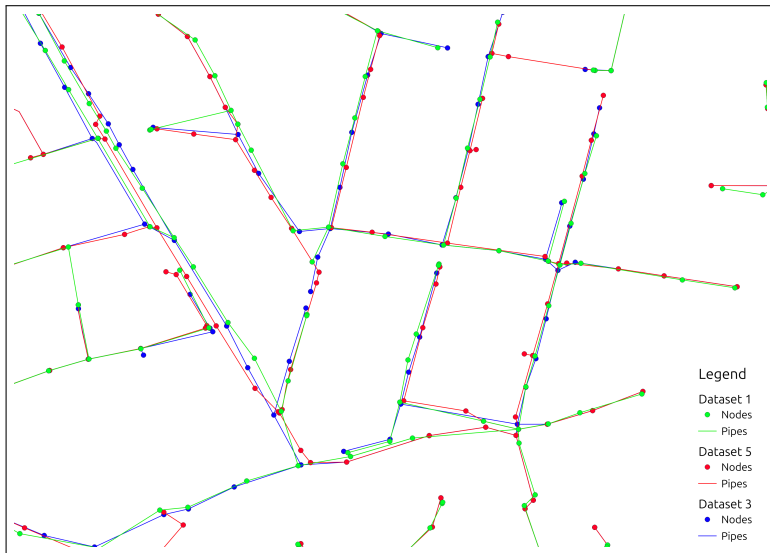
Autres formes de données/informations hétérogènes

- Données textuelles, contraintes, etc.
- Données imparfaites : non fiables, imprécises, incomplètes, dynamiques, conflictuelles, etc.

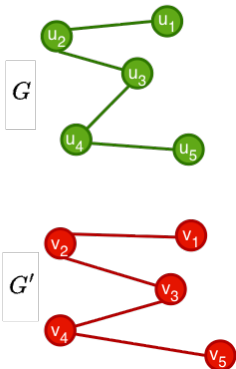
Comment représenter ces données hétérogènes multi-sources?



Trois jeux de données hétérogènes d'assainissement de Montpellier



Input Networks



Définition

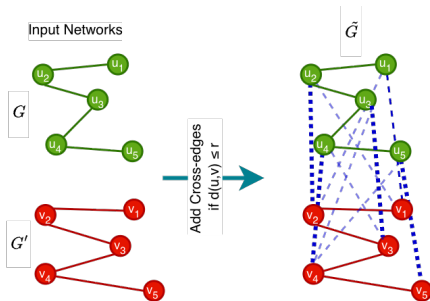
Étant donnés deux réseaux $G = (V, E)$ et $G' = (V', E')$ représentant la **même infrastructure** depuis des sources différentes, trouver une application :

$$\pi^* : V \rightarrow V'$$

qui identifie les **nœuds correspondants** entre les réseaux. **Dans notre contexte**

- Nœuds = regards, jonctions
- Arêtes = conduites
- Objectif : trouver **toutes** les correspondances

Étape 1 : Graphe composé



Graphe composé :

$$\tilde{G} = (V \cup V', E \cup E' \cup E_{\text{cross}})$$

Filtrage spatial :

$$E_{\text{cross}} = \{(u, v) \mid u \in V, v \in V', d(u, v) \leq r\}$$

où $d(u, v)$ est la distance euclidienne, et r est l'hyperparamètre de rayon spatial.

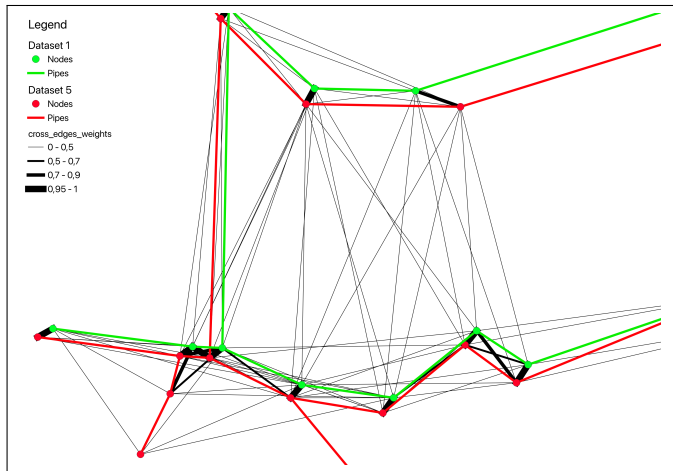
Poids des arêtes inter-graphes (décroissance gaussienne) :

$$w(u, v) = e^{-d(u,v)^2 / (2\sigma^2)}$$

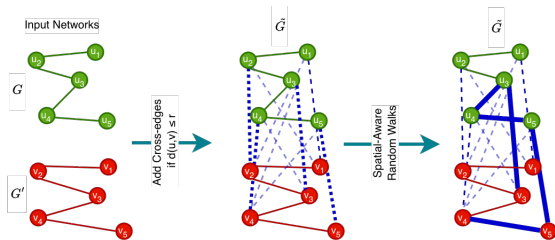
Complexité : $O(n \log n + k(r) \cdot n) \ll O(n^2)$
où $n = \max(|V|, |V'|)$ et $k(r)$ est le nombre maximal de candidats spatiaux par nœud.

Étape 1 : Graphe composé

Graphe composé montrant des arêtes inter-graphes.



Étape 2 : Marche aléatoire tenant compte de l'information spatiale



Transition de marche aléatoire :

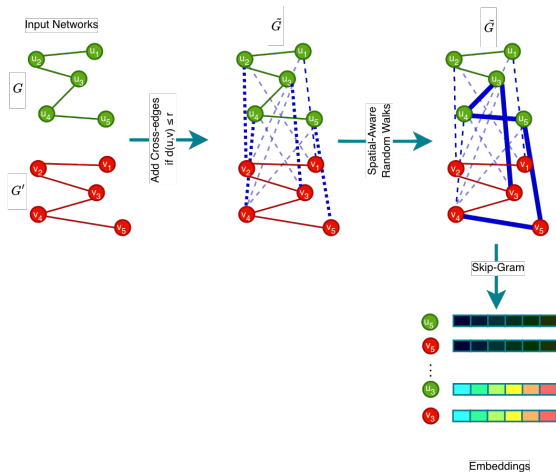
$$P(v_{t+1}|v_t) = \frac{w(v_t, v_{t+1})}{\sum_{u \in N(v_t)} w(v_t, u)}$$

Exemple de marche sur $\tilde{G}$:

$$v_4 \xrightarrow{w=1} v_5 \xrightarrow{w=0.9} u_5 \xrightarrow{w=1} u_4 \xrightarrow{w=1} u_3 \xrightarrow{w=0.8} v_3$$

Arêtes bleues : sauts inter-graphes avec probabilité $\propto$ à la proximité spatiale.

Étape 4 : Apprentissage des embeddings



Objectif Skip-Gram :

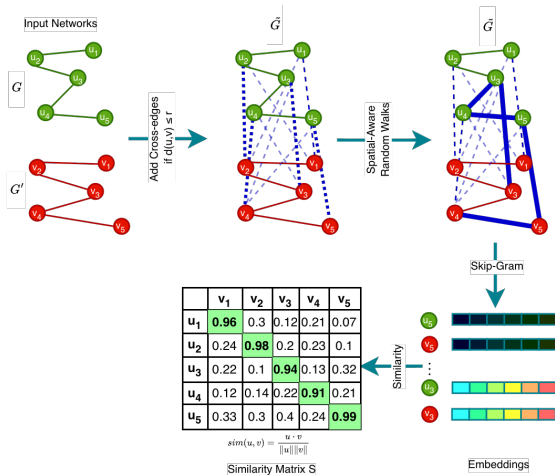
$$\mathcal{L}_{SG} = \max \sum_{v_i \in \text{walks}} \sum_{v_j \in \text{context}(v_i)} \log P(v_j | v_i)$$

Échantillonnage négatif :

$$\mathcal{L}_{NS} = \log \sigma(x'_{v_j}{}^T x_{v_i}) + \sum_{k=1}^K \mathbb{E}_{v_n \sim P_n(v)} [\log \sigma(-x'_{v_n}{}^T x_{v_i})]$$

Idée clé : des nœuds de réseaux différents mais spatialement proches co-apparaissent dans les marches $\Rightarrow$ embeddings similaires.

Étape 4 : Calcul de similarité



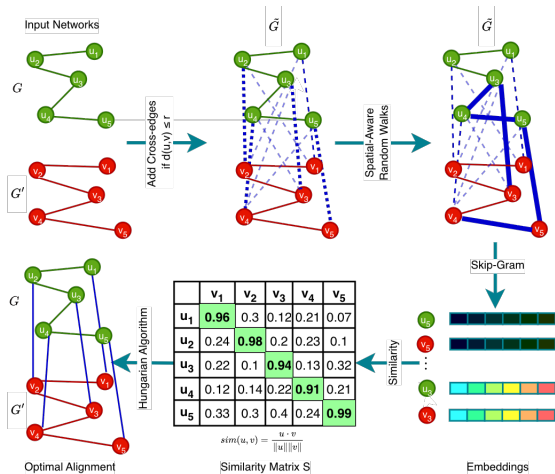
Similarité cosinus :

$$sim(u, v) = \frac{x_u \cdot x_v}{\|x_u\| \|x_v\|}$$

Calculée uniquement pour les candidats spatiaux $v \in V'$ tels que $d(u, v) \leq r$.

Idée clé : des nœuds co-localisés dans différents réseaux acquièrent des embeddings similaires.

Étape 5 : Affectation optimale



Matrice de coût :

$$C_{ij} = 1 - \text{sim}(u_i, v_j)$$

Affectation optimale :

$$\pi^* = \arg \min_{\pi} \sum_{u \in V} C(u, \pi(u))$$

où $\pi : V \rightarrow V'$ est une application bijective.

Résultat : alignement globalement optimal un-à-un maximisant la similarité totale.

Réseau d'assainissement — Montpellier

	D1	D5
V (Complet)	43,826	55,027
E (Complet)	44,200	55,600
V (Prades)	805	977
E (Prades)	816	981

Défi : degré médian = 2, structure arborescente
⇒ nœuds structurellement identiques ⇒
méthodes structurelles échouent.

Performance d'alignement (F1-Score %)

Méthode	Complet	Prades
SANA	95.7±0.1	95.8±0.1
CENA	Infaisable	4.76
Struc2vec	Infaisable	27.54
DeepWalk	1.28	30.2

La spatialité est essentielle pour un alignement scalable.

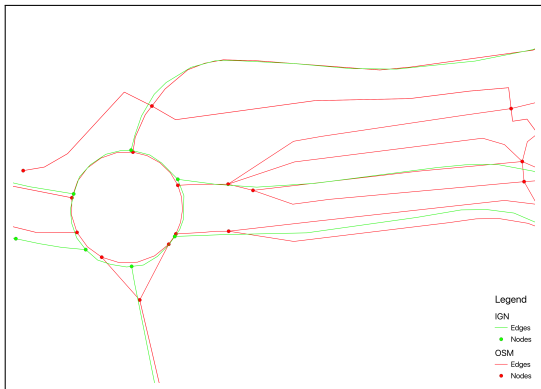


Figure: Désalignement spatial entre les réseaux routiers IGN et OSM.

Jeu de données (Bouches-du-Rhône, France) :

- IGN : 388,148 nœuds
- OSM : 250,302 nœuds
- 9× plus grand que le jeu de données d'assainissement
- Vérité terrain : 8,531 paires (seuil 0.5m)

Résultats SANA :	Précision	96.07%
	Rappel	94.76%
	F1-Score	95.41%

8,084 sur 8,531 ancres correctement identifiées.

Contributions de SANA :

- Graphe composé avec arêtes gaussiennes
- $O(n \log n)$ vs $O(n^2)$
- Sur graphes homogènes, les mesures de similarité structurelle échouent $\Rightarrow$ la spatialité est indispensable

Perspectives :

- Attributs sémantiques & temporalité
- Validation sur réseaux électriques & télécoms
- Intégration multi-opérateurs