Anonymisation et Compression de Graphes

Alexis Guyot

⟨alexis_guyot@etu.u-bourgogne.fr⟩

M2 BDIA
Université de Bourgogne

01/04/2021

Sommaire

Sommaire

Introduction à l'anonymisation de graphes

Contexte

- Réseaux sociaux, sources de données en constante expansion.
  - Utilisés par 51% de la population mondiale.
  - Croissance moyenne de +10% entre 2019 et 2020.

Introduction à l'anonymisation de graphes

Contexte

- Réseaux sociaux, sources de données en constante expansion.
  - Utilisés par 51% de la population mondiale.
  - Croissance moyenne de +10% entre 2019 et 2020.
- Données oui, mais données privées.
  - Règlement Général sur la Protection des Données (RGPD) : "Toute information se rapportant à une personne identifiée ou identifiable".

Introduction à l'anonymisation de graphes

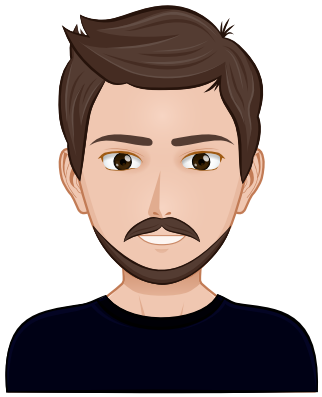
Contexte

- Réseaux sociaux, sources de données en constante expansion.
  - Utilisés par 51% de la population mondiale.
  - Croissance moyenne de +10% entre 2019 et 2020.
- Données oui, mais données privées.
  - Règlement Général sur la Protection des Données (RGPD) : "Toute information se rapportant à une personne identifiée ou identifiable".
- Comment publier ces données en préservant l'anonymat des utilisateurs ?

Introduction à l'anonymisation de graphes

Principales failles de sécurité

- Découverte d'identité (*Identity disclosure*).

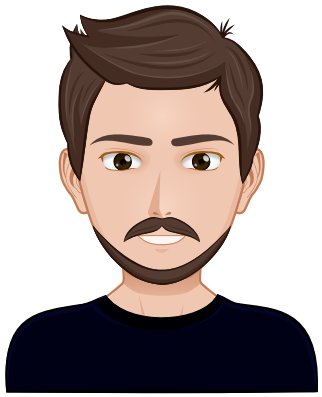


"Je sais que le sommet 0 correspond à Barack Obama !"

Introduction à l'anonymisation de graphes

Principales failles de sécurité

- Découverte d'identité (*Identity disclosure*).
- Découverte d'attribut (*Content disclosure*).

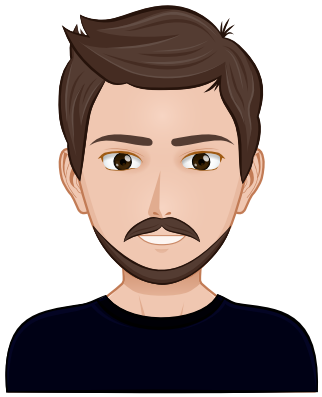


"Je sais que le sommet 0 est un homme !"

Introduction à l'anonymisation de graphes

Principales failles de sécurité

- Découverte d'identité (*Identity disclosure*).
- Découverte d'attribut (*Content disclosure*).
- Découverte de lien (*Link disclosure*).



"Je sais que les sommets 0 et 1 sont liés !"

ou

"Je sais que le poids de la relation entre les sommets 0 et 1 vaut 3 !"

k-Anonymat

k-Anonymat

Au moins k sommets possèdent les caractéristiques Q.

k-Anonymat

k-Anonymat

Au moins k sommets possèdent les caractéristiques Q.

- Q : Information sur le degré, sur le voisinage, ...

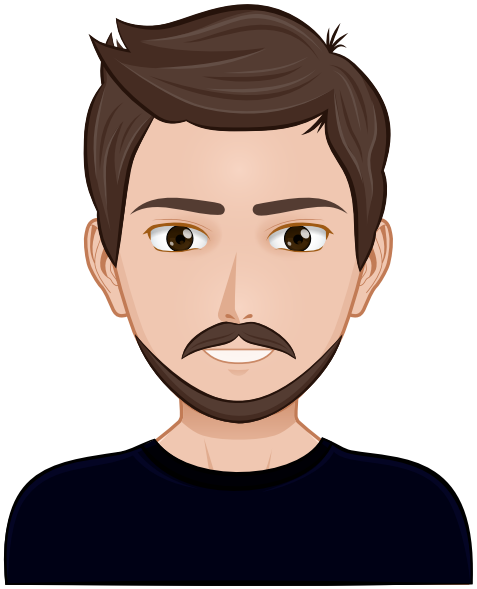
k-Anonymat

k-Anonymat

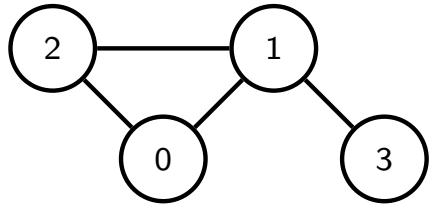
Au moins k sommets possèdent les caractéristiques Q.

- Q : Information sur le degré, sur le voisinage, ...
- Chaque sommet est indissociable de $k - 1$ autres.

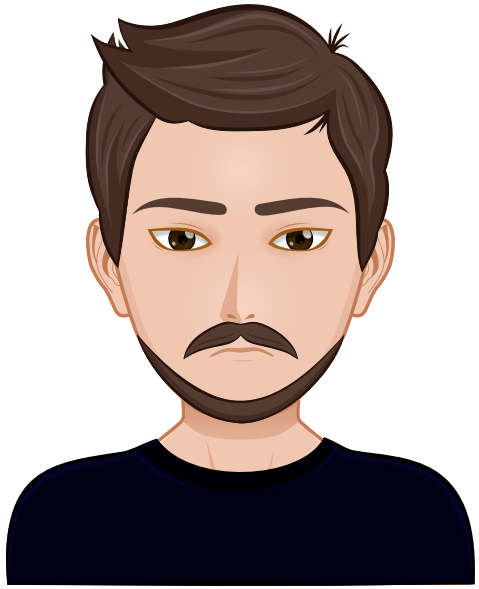
k-Anonymat



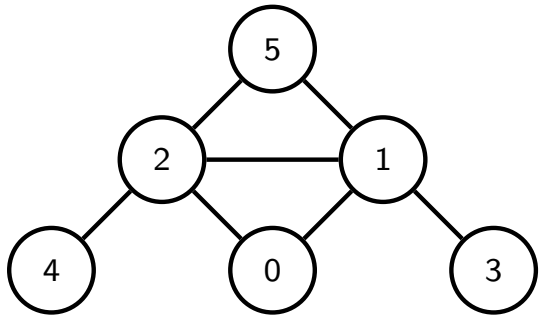
"Je connais un utilisateur relié à un triangle."



k-Anonymat



"Je connais un utilisateur relié à un triangle."



Privacy/Utility Tradeoff

Comment publier des graphes constitués de données privées en préservant à la fois leur utilité et l'anonymat des personnes représentées ?

Familles de méthodes

Anonymiser par généralisation

- (SKARKALA et al., 2012)
- (CASAS-ROMA ; ROUSSEAU, 2015)
- (CAMPAN ; TRUTA, 2008)
- (HAY et al., 2008)
- (BONCHI ; GIONIS ; TASSA, 2014)

Familles de méthodes

Anonymiser par généralisation

- (SKARKALA et al., 2012)
- (CASAS-ROMA ; ROUSSEAU, 2015)
- (CAMPAN ; TRUTA, 2008)
- (HAY et al., 2008)
- (BONCHI ; GIONIS ; TASSA, 2014)

Par ajout de bruit déterministe

- (FEDER ; NABAR ; TERZI, 2008)

Familles de méthodes

Anonymiser par généralisation

- (SKARKALA et al., 2012)
- (CASAS-ROMA ; ROUSSEAU, 2015)
- (CAMPAN ; TRUTA, 2008)
- (HAY et al., 2008)
- (BONCHI ; GIONIS ; TASSA, 2014)

Par ajout de bruit déterministe

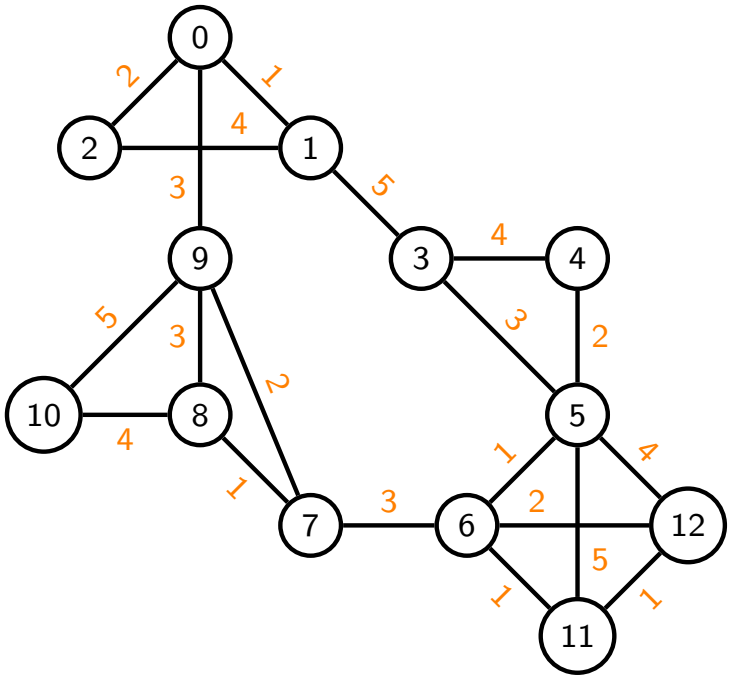
- (FEDER ; NABAR ; TERZI, 2008)

Par ajout de bruit probabiliste

- (BOLDI et al., 2012)
- (NGUYEN, 2016)
- (BONCHI ; GIONIS ; TASSA, 2014)

Sommaire

Anonymisation par généralisation

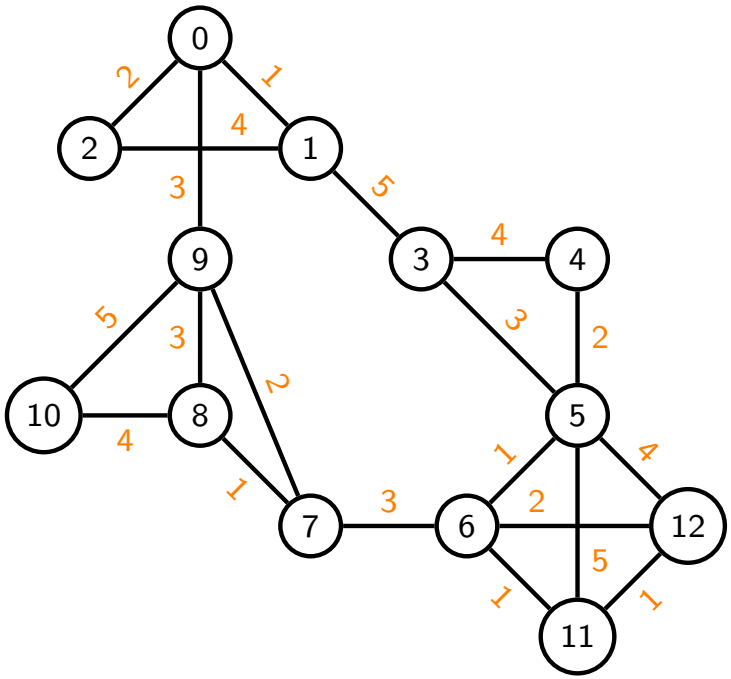


Principe général

- Regrouper en super-nœuds et super-arêtes.
- Résumer le graphe d'origine dans les attributs.

Résumer la structure du graphe : Casas et Rousseau

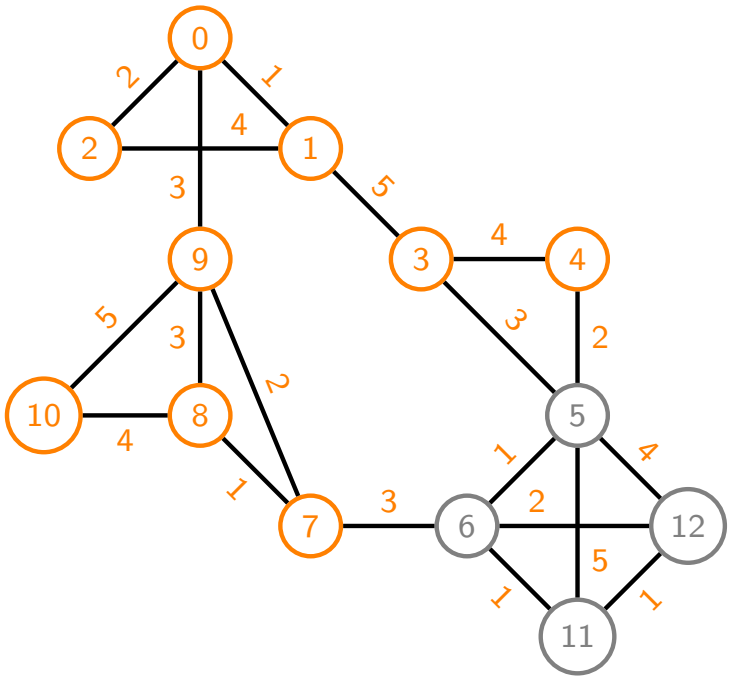
Regroupement en super-nœuds.



Résumer la structure du graphe : Casas et Rousseau

Regroupement en super-nœuds.

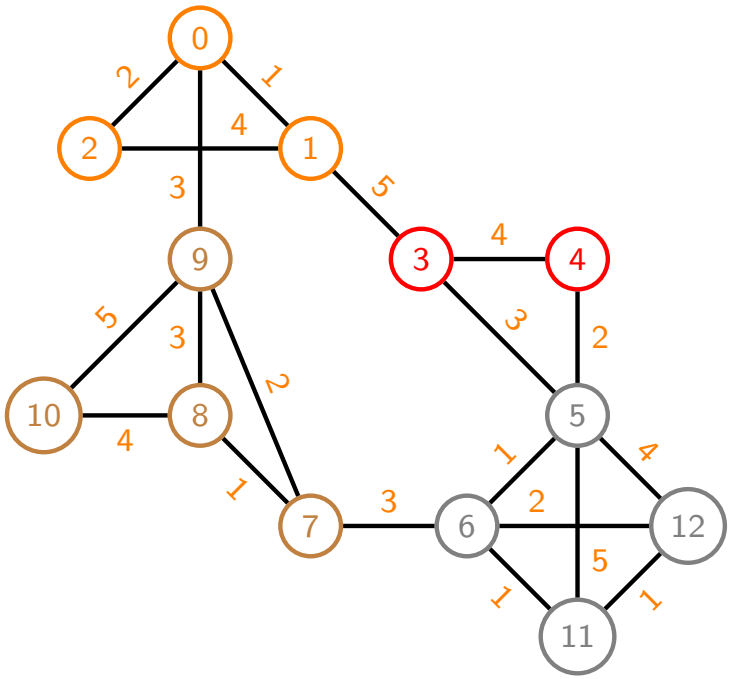
- Classification en k-shells.

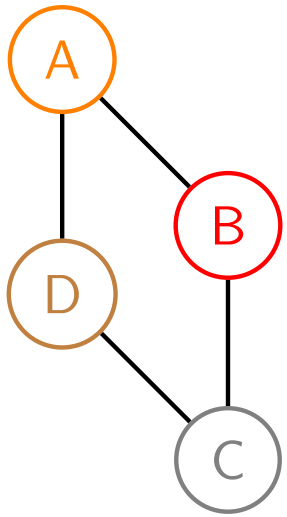


Résumer la structure du graphe : Casas et Rousseau

Regroupement en super-nœuds.

- Classification en k-shells.
- *Multilevel/FastGreedy*, *Manhattan/2-path*.

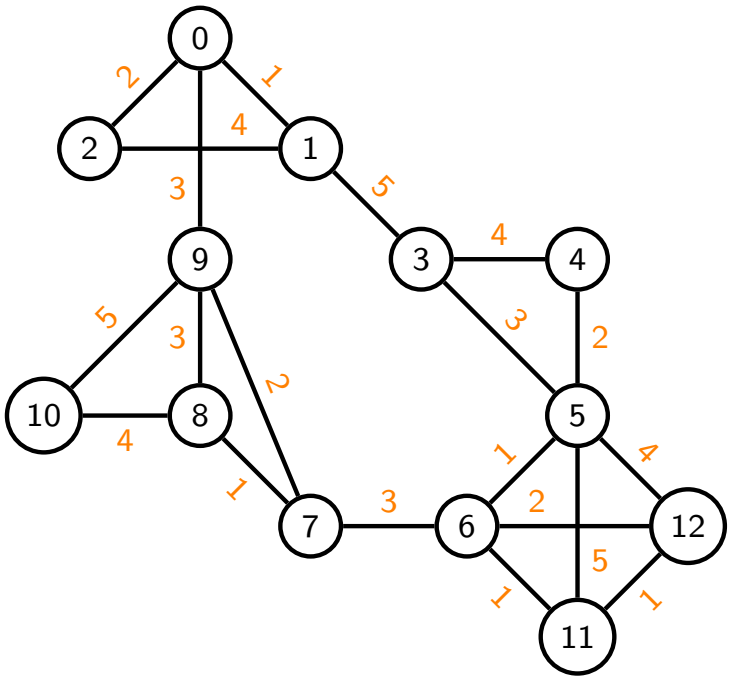


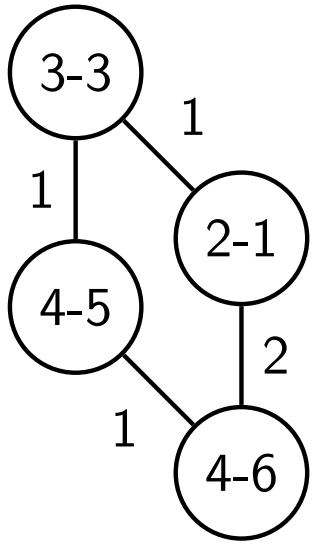


Résumer la structure du graphe : Casas et Rousseau

Résumer la structure d'origine.

- Informations intra-nœuds dans les attributs.
- Informations inter-nœuds dans la pondération.





Pondération et découverte de lien : Skarkala et al.

Regroupement en super-nœuds (respecte k-anonymat).

- Regrouper au hasard.

Pondération et découverte de lien : Skarkala et al.

Regroupement en super-nœuds (respecte k-anonymat).

- Regrouper au hasard.
- Regroupement qui minimise la perte d'information.
  - G = (V,E) un graphe et G'=(V',E') sa version anonymisée.
  - W(e) le poids de l'arête e.
  - Perte = $\frac{1}{|E|}\sum_{e\in E}(W(e) - W(e'))^2$, où $e \in e'$ dans E'.

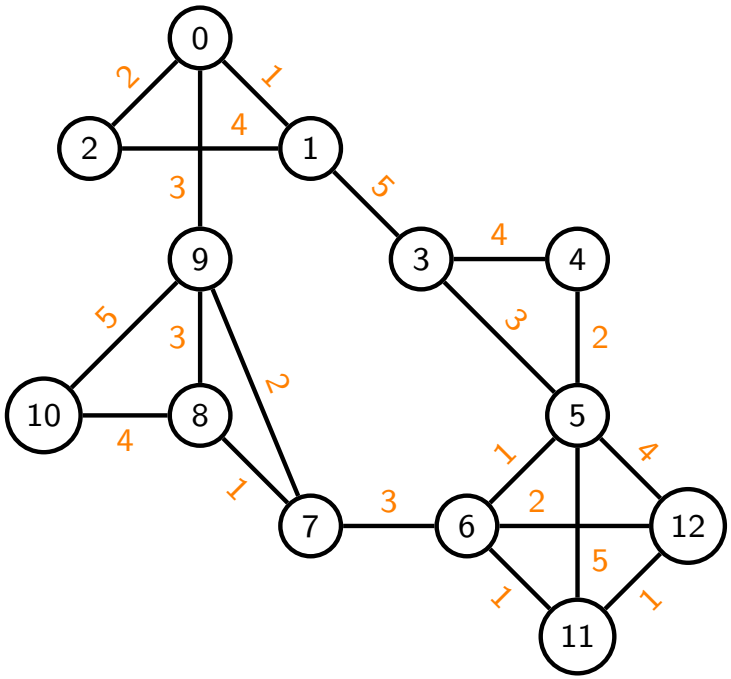
Pondération et découverte de lien : Skarkala et al.

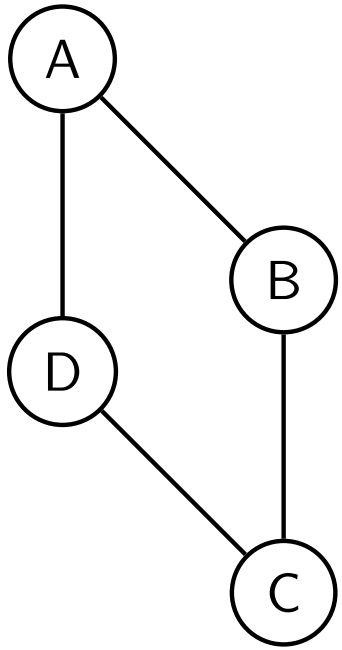
Regroupement en super-nœuds (respecte k-anonymat).

- Regrouper au hasard.
- Regroupement qui minimise la perte d'information.
  - G = (V,E) un graphe et G'=(V',E') sa version anonymisée.
  - W(e) le poids de l'arête e.
  - Perte = $\frac{1}{|E|}\sum_{e\in E}(W(e)-W(e'))^2$, où $e \in e'$ dans E'.
- Regroupement qui n'augmente pas trop la perte d'information (seuil).

Pondération et découverte de lien : Skarkala et al.

Résumer la structure d'origine.

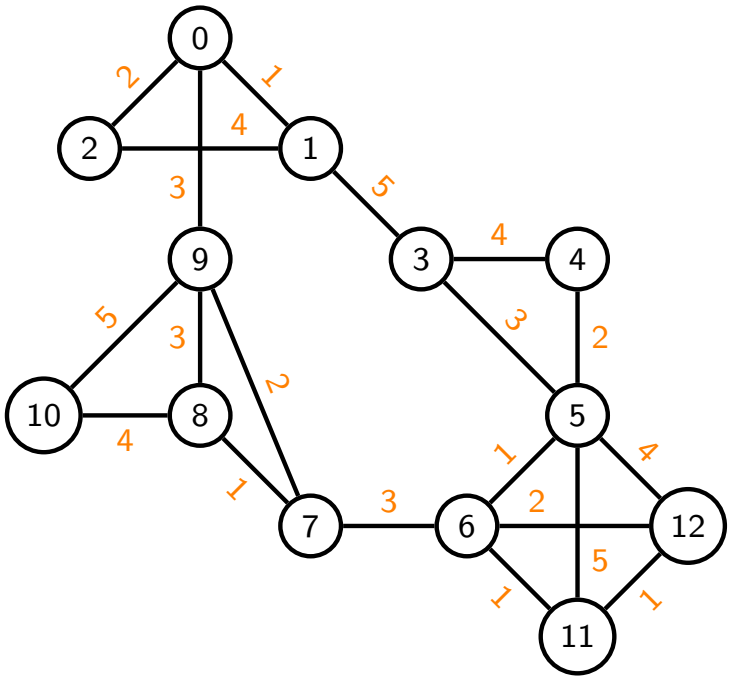


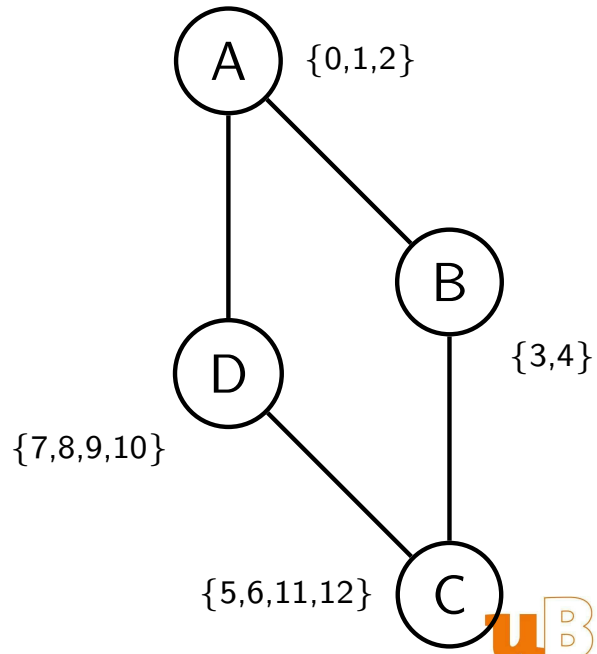


Pondération et découverte de lien : Skarkala et al.

Résumer la structure d'origine.

- Composition des super-nœuds en attributs.

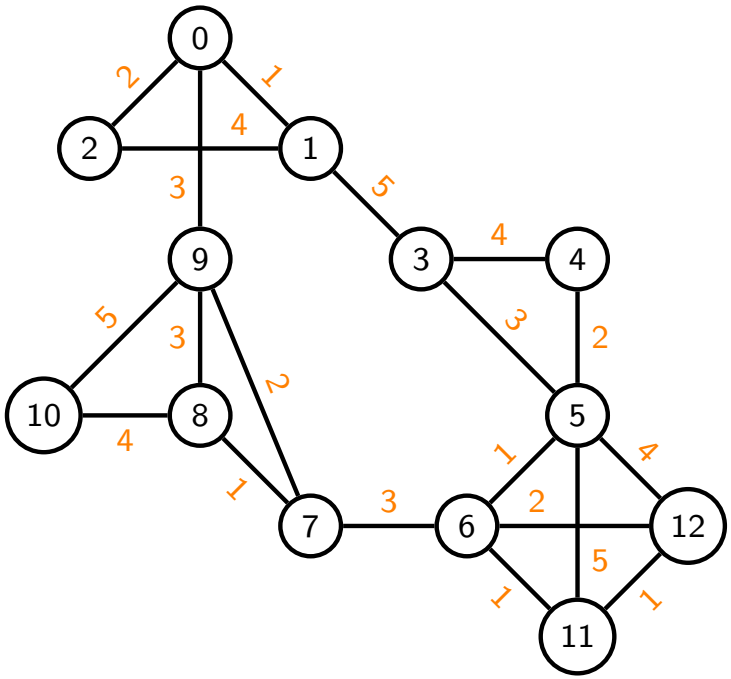


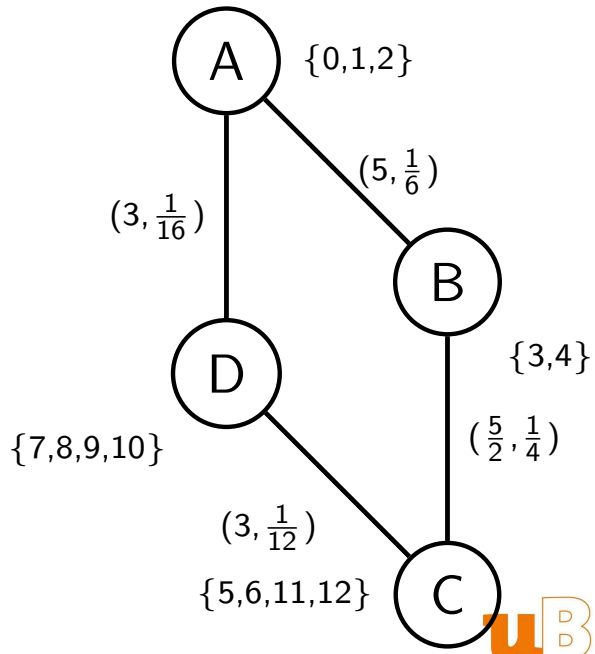


Pondération et découverte de lien : Skarkala et al.

Résumer la structure d'origine.

- Composition des super-nœuds en attributs.
- Pondération en deux parties : (moyenne des poids, probabilité).

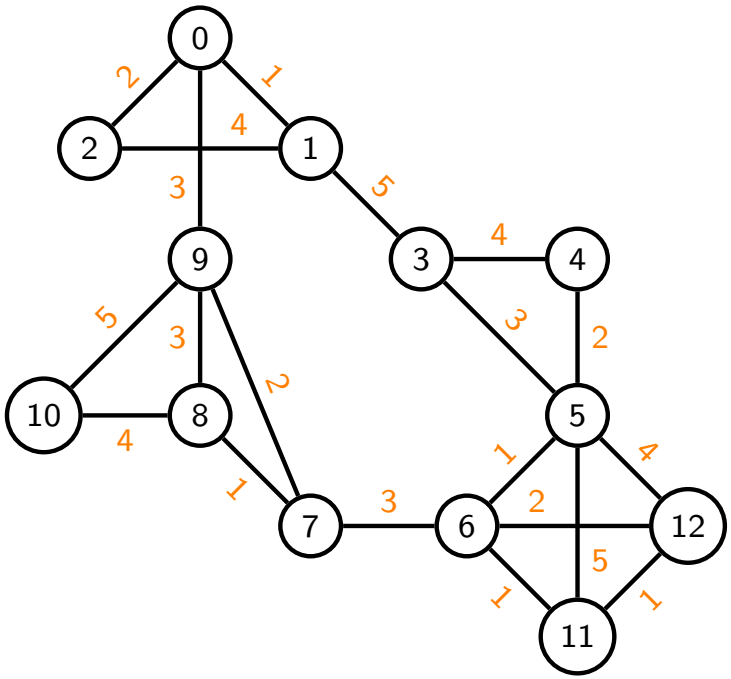


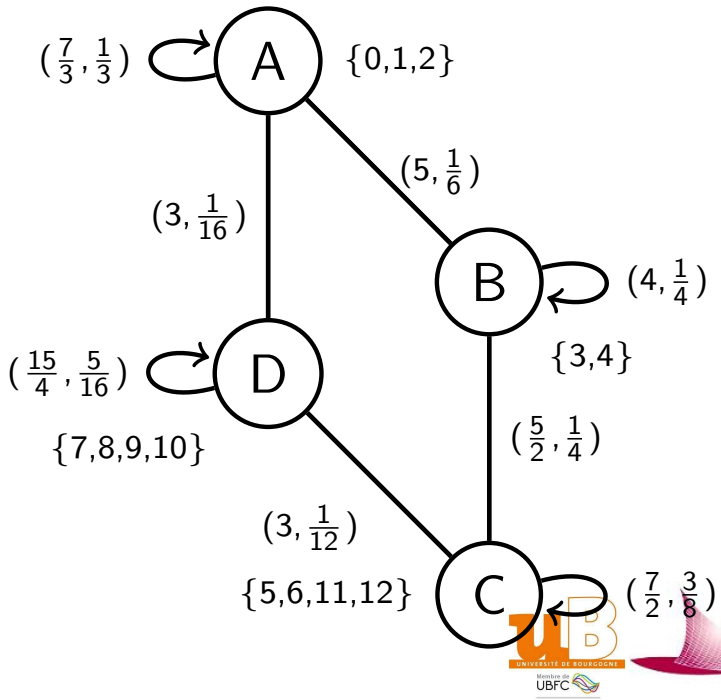


Pondération et découverte de lien : Skarkala et al.

Résumer la structure d'origine.

- Composition des super-nœuds en attributs.
- Pondération en deux parties : (moyenne des poids, probabilité).





Généraliser les attributs : Campan et Truta

Généraliser la structure (respecte k-anonymat).

Généraliser les attributs : Campan et Truta

Généraliser la structure (respecte k-anonymat).

- Deux mesures pour quantifier la perte d'information.

Généraliser les attributs : Campan et Truta

Généraliser la structure (respecte k-anonymat).

- Deux mesures pour quantifier la perte d'information.
- Algorithme glouton qui minimise les deux valeurs.

Généraliser les attributs : Campan et Truta

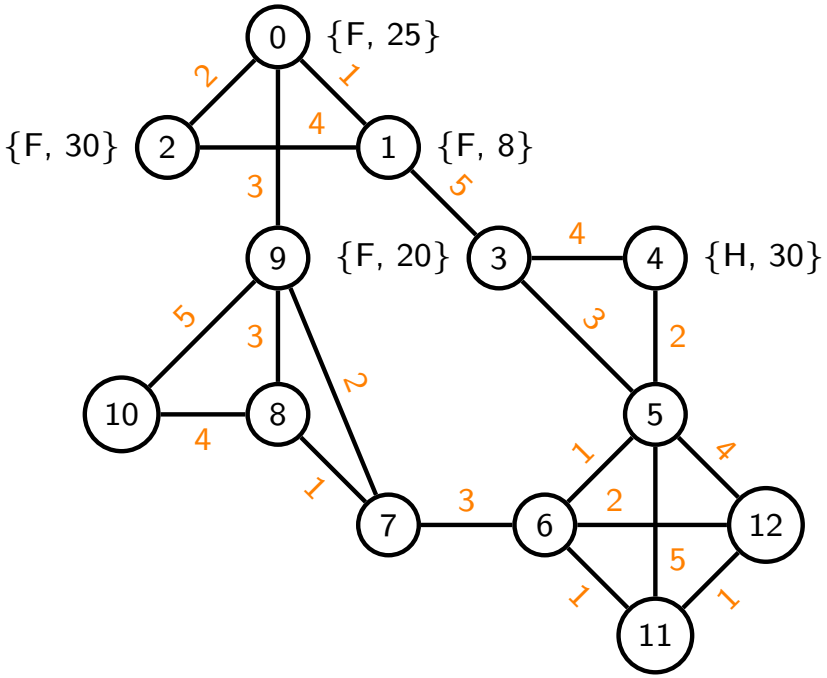
Généraliser la structure (respecte k-anonymat).

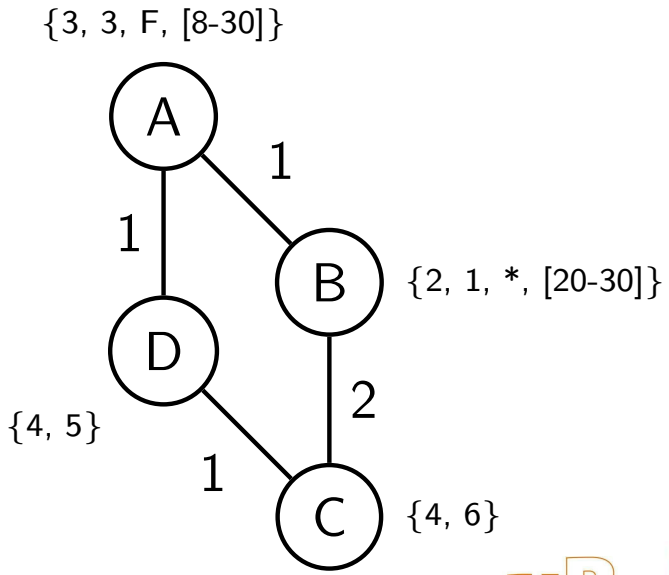
- Deux mesures pour quantifier la perte d'information.
- Algorithme glouton qui minimise les deux valeurs.
- Pondération des super-arêtes : nombre d'arêtes résumées.

Généraliser les attributs : Campan et Truta

Généraliser les attributs.

- Attributs numériques : intervalle de valeurs.
- Attributs catégories : catégorie parente (concrète ou abstraite).





Conclusion sur la généralisation

- Avantages
  - ⊕ Couvre naturellement la découverte d'identité et de liens.

Conclusion sur la généralisation

- Avantages
  - ⊕ Couvre naturellement la découverte d'identité et de liens.
  - ⊕ Quelques pistes contre la découverte d'attributs.

Conclusion sur la généralisation

- Avantages
  - ⊕ Couvre naturellement la découverte d'identité et de liens.
  - ⊕ Quelques pistes contre la découverte d'attributs.
  - ⊕ Information supprimée mais résumée.

Conclusion sur la généralisation

- Avantages
  - ⊕ Couvre naturellement la découverte d'identité et de liens.
  - ⊕ Quelques pistes contre la découverte d'attributs.
  - ⊕ Information supprimée mais résumée.
  - ⊕ Réduction de la taille du graphe.

Conclusion sur la généralisation

- Avantages
  - ⊕ Couvre naturellement la découverte d'identité et de liens.
  - ⊕ Quelques pistes contre la découverte d'attributs.
  - ⊕ Information supprimée mais résumée.
  - ⊕ Réduction de la taille du graphe.
- Inconvénient
  - ⊖ Graphes en sortie très différents des graphes en entrées.

Sommaire

Anonymisation par ajout de bruit

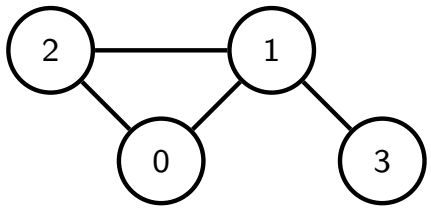
Deux approches :

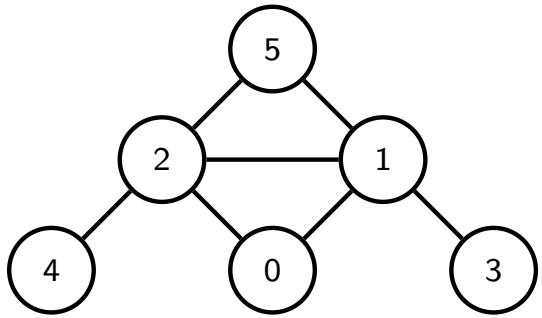
- Bruit déterministe : Ensemble de règles à appliquer pour ajouter des informations jusqu'à atteindre un certain niveau d'anonymat.
- Bruit probabiliste : Ajouter/Retirer/Échanger des données de manière aléatoire ou transformer la pondération en probabilités.

Ajout de bruit déterministe

Principe général.

- Créer pertinemment de fausses données.
- Ajouter des sommets et arêtes jusqu'à atteindre un niveau de sécurité souhaité.

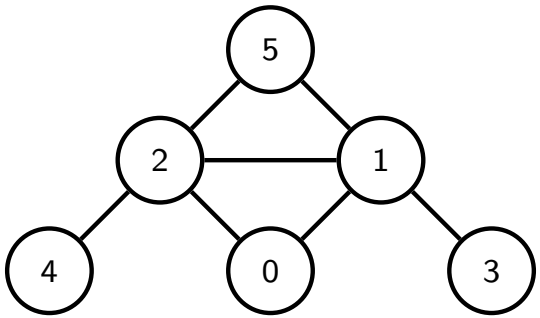




Ajout de bruit déterministe : Feder et al.

Créer pertinemment de fausses données.

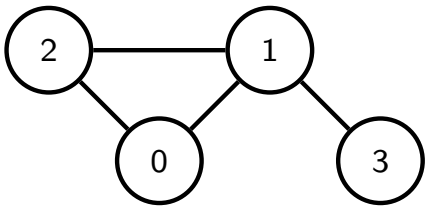
- Spécialisée pour le k-anonymat de voisinage (k-voisinage).
- Cherche le minimum d'arêtes à ajouter pour atteindre le (k,l)-anonymat.
- Exploite les spécificités apportées par certaines valeurs de k et de l.

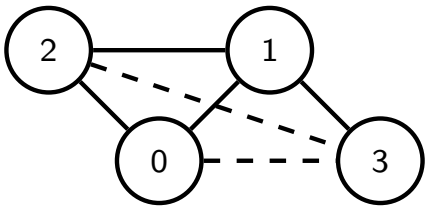


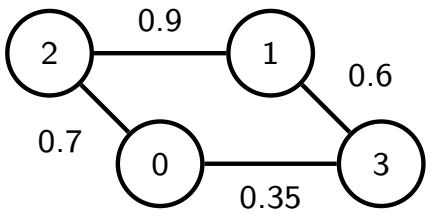
Ajout de bruit probabiliste : Boldi et al.

Créer un ensemble de mondes possibles.

- La pondération de chaque arête devient une probabilité d'exister.
- Travaille sur l'entièreté des arêtes possibles dans un graphe.







Ajout de bruit probabiliste : Boldi et al.

k-offuscation.

- Entropie de Shannon pour évaluer la protection apportée par la perturbation.

Ajout de bruit probabiliste : Boldi et al.

k-offuscation.

- Entropie de Shannon pour évaluer la protection apportée par la perturbation.
- Probabilité de correctement découvrir l'identité d'un sommet à partir d'une information structurelle.

Ajout de bruit probabiliste : Boldi et al.

Grandes lignes de la méthode.

- Calcul d'un score d'unicité pour chaque sommet.

Ajout de bruit probabiliste : Boldi et al.

Grandes lignes de la méthode.

- Calcul d'un score d'unicité pour chaque sommet.
- Des perturbations sont générées à partir des distributions normale et uniforme.

Ajout de bruit probabiliste : Boldi et al.

Grandes lignes de la méthode.

- Calcul d'un score d'unicité pour chaque sommet.
- Des perturbations sont générées à partir des distributions normale et uniforme.
- Plus un sommet est unique, plus ses arêtes incidentes recevront une forte perturbation.

Ajout de bruit probabiliste : Boldi et al.

Grandes lignes de la méthode.

- Calcul d'un score d'unicité pour chaque sommet.
- Des perturbations sont générées à partir des distributions normale et uniforme.
- Plus un sommet est unique, plus ses arêtes incidentes recevront une forte perturbation.
- k-offuscation évaluée puis le processus est réitéré.

Conclusion sur l'ajout de bruit

- Avantages
  - ⊕ Structure finale proche de celle d'origine.

Conclusion sur l'ajout de bruit

- Avantages
  - ⊕ Structure finale proche de celle d'origine.
  - ⊕ Bonne préservation de certaines caractéristiques globales (distribution des degrés, diamètre, etc.)

Conclusion sur l'ajout de bruit

- Avantages
  - ⊕ Structure finale proche de celle d'origine.
  - ⊕ Bonne préservation de certaines caractéristiques globales (distribution des degrés, diamètre, etc.)
  - ⊕ Pour les méthodes probabilistes, crée tout un ensemble de solutions possibles.

Conclusion sur l'ajout de bruit

- Avantages
  - ⊕ Structure finale proche de celle d'origine.
  - ⊕ Bonne préservation de certaines caractéristiques globales (distribution des degrés, diamètre, etc.)
  - ⊕ Pour les méthodes probabilistes, crée tout un ensemble de solutions possibles.
- Inconvénients
  - ⊖ Difficultés à préserver les structures locales, comme les communautés et les cliques.

Conclusion sur l'ajout de bruit

- Avantages
  - ⊕ Structure finale proche de celle d'origine.
  - ⊕ Bonne préservation de certaines caractéristiques globales (distribution des degrés, diamètre, etc.)
  - ⊕ Pour les méthodes probabilistes, crée tout un ensemble de solutions possibles.
- Inconvénients
  - ⊖ Difficultés à préserver les structures locales, comme les communautés et les cliques.
  - ⊖ Peu de solutions pour la découverte d'attributs.

Sommaire

Conclusion

- Modifier un graphe pour ajouter des garanties de k-anonymat tout en préservant l'utilité des données.

Conclusion

- Modifier un graphe pour ajouter des garanties de k-anonymat tout en préservant l'utilité des données.
- Deux grandes familles de méthodes :
  - par généralisation : compression avec pertes.
  - par ajout de bruit : déterministe ou probabiliste.

Conclusion

- Modifier un graphe pour ajouter des garanties de k-anonymat tout en préservant l'utilité des données.
- Deux grandes familles de méthodes :
  - par généralisation : compression avec pertes.
  - par ajout de bruit : déterministe ou probabiliste.
- Quel rôle pour les méthodes de compression sans pertes ?

Bibliographie


BOLDI, P. et al. Injecting uncertainty in graphs for identity obfuscation. *arXiv preprint arXiv :1208.4145*, 2012.


BONCHI, F. ; GIONIS, A. ; TASSA, T. Identity obfuscation in graphs through the information theoretic lens. *Information Sciences*, Elsevier, v. 275, p. 232–256, 2014.


CAMPAN, A. ; TRUTA, T. M. *A clustering approach for data and structural anonymity in social networks*. [S.l.] : PinKDD, 2008.


CASAS-ROMA, J. ; ROUSSEAU, F. Community-preserving generalization of social networks. In : IEEE. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.], 2015. p. 1465–1472.


FEDER, T. ; NABAR, S. U. ; TERZI, E. Anonymizing graphs. *arXiv preprint arXiv :0810.5578*, 2008.


HAY, M. et al. Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 1, n. 1, p. 102–114, 2008.


NGUYEN, H.-H. *Social Graph Anonymization*. Tese (Doutorado), 2016.


SKARKALA, M. E. et al. Privacy preservation by k-anonymization of weighted social networks. In : IEEE. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. [S.l.], 2012. p. 423–428.

Remerciements

Merci pour votre attention. Des questions ?