TIPE : Etude de couvertures de réseaux de métros, application de l'homologie persistante

Elowan
elowarp@gmail.com

Contents

Abstract

Nous nous proposons ici d'étudier les différentes disparités dans les réseaux métropolitain de plusieurs grandes villes, dans le sens où l'on veut détecter les zones spatiales les plus en besoin de dévéloppement de transports en commun. Cela grâce une approche utilisant de l'analyse topologique, l'homologie persistante, qui se veut dans notre cas être une approche plus pertinente que l'étude des distances spaciales entre les stations de métros.

I Définitions

Commençons par définir l'homologie persistante intuitivement :

> **Définition :**
> L'homologie persistante est une méthode d'analyse de données topologiques. Elle est capable de donner une caractérisation de la naissance de trous à de multiples dimensions.

L'exemple suivant est tiré de [1].

L'homologie persistante essaye de formaliser le processus que le cerveau a pour interpréter les objets. Prenons comme exemple le style artistique du poitillisme.

En effet, lorsque l'on regarde un tableau de Seaurat (Figure 1), nous sommes capable d'en comprendre le contexte, de reconnaitre les objets. Cela n'est pas vrai que pour les tableaux de Seurat mais aussi pour les images pixélisées ou encore avec la paréidolie faciale (le fait de voir des visages là où il n'y en a pas normalement). L'homologie persistante essaye alors de donner cette description depuis un ensemble discret de points.

Pour l'utilisation que nous allons en faire dans cette étude, nous utiliserons l'homologie persistante afin de detecter les "trous" dans un ensemble discret de point représentant des stations de métros.

Figure 1: La scène à la Grande Jatte - Printemps (Georges Seurat, 1888)

I.1 Constructions géométriques

Voyons désormais comment formaliser cela, commençons par définir tous nos outils :

Définition :
Un *simplexe* est l'analogue du triangle à k dimensions, c'est l'objet le plus simple qu'il est possible de définir en k dimensions.

Par exemple, un simplexe en dimension 0 est un point, en dimension 1 c'est une droite, en dimension 3 c'est une pyramide et ainsi de suite.

Remarque : On dit que σ_i est une face de σ_j si σ_j fait parti des bords du simplexe défini par σ_i, donc σ_i est nécessairement de dimension supérieure de 1 à σ_j

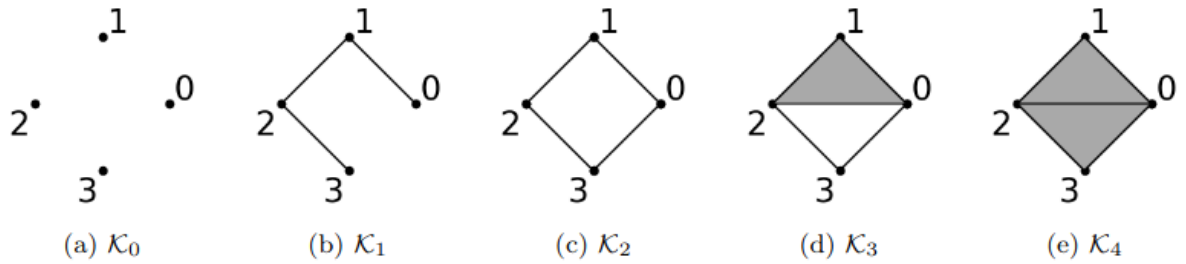
On a donc naturellement la définition suivante :

Définition :
Un *complexe simplicial* est un ensemble de simplexes. On dit que le complexe est de dimension k si la dimension maximale de ses simplexes est k.

On donne une représentation de plusieurs complexes simplicaux Figure 2.

Afin de pouvoir caractériser des changements, il faut pouvoir définir deux états différents à comparer, une filtration permet alors d'ordonner ces différents états de façon à en étudier les changements.

Définition :
Une *filtration* est une application qui à un entier i associe un complexe simplicial K_i de telle sorte que $\forall j \in [\![0, i]\!]$ le complexe simplicial K_j est inclus dans K_i



(a) $\mathcal{K}_0$ (b) $\mathcal{K}_1$ (c) $\mathcal{K}_2$ (d) $\mathcal{K}_3$ (e) $\mathcal{K}_4$

Figure 2: Représentation d'une filtration où $K_0 \subset K_1 \subset K_2 \subset K_3 \subset K_4$, tiré de [2]

Grâce à cette definition, nous sommes capable de quantifier les changements d'un complexe à l'autre, comme la création de cycles (dans K_2, il apparait un cycle $(0, 1, 2, 3)$ après avoir rajouté $(0, 3)$ à K_1) ou la destruction de composantes connexes (dans K_0 tous les simplexes 0D sont dans des composantes connexes différentes alors que dans K_1 ils sont tous dans la même, on a "cassé" les composantes connexes de 1, 2 et 3).

Le problème que nous avons est que nous voulons analyser un ensemble de point discret, et non pas une filtration déjà existante, il nous faut alors créer une filtration depuis un ensemble de points. Cela va se faire via une construction incrémentale de complexes simplicials, par soucis d'implémentation, nous choisissons comme dans [2] les complexes de Vietoris-Rips pondérés :

Définition :
Soient un ensemble $X = (x_i)_{i=0}^n$ de points de poids $(w_i)_{i=0}^n$ et une distance d, on définit le complexe simplicial pondéré Vietoris-Rips au rang r $V_r^w(X, \mathbb{R}^2, d)$ comme l'ensemble des simplexes $\left(x_{i_0}, ..., x_{i_k}\right)$ tels que :

- $\forall j \in [\![0, k]\!], w_{i_k} < r$
- $\forall (j, l) \in [\![0, k]\!]^2, d\left(x_{i_j}, x_{i_k}\right) + w_{i_j} + w_{i_k} < 2t$

Ainsi plus on augmente r, plus le complexe possède des simplexes, on en donne une représentation Figure 3.

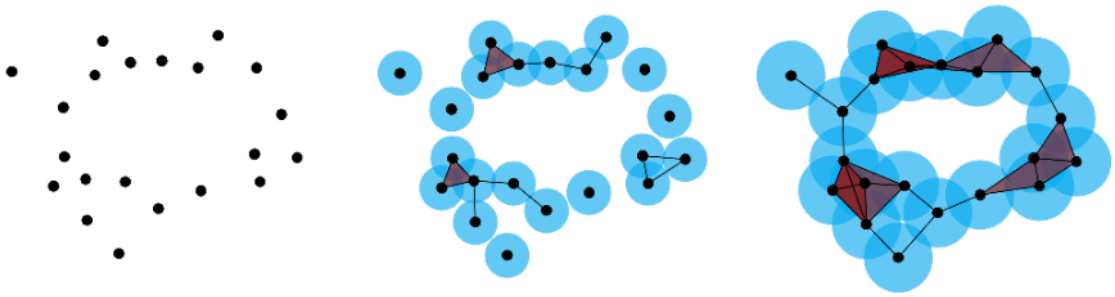


Figure 3: Construction d'un complexe simplicial avec un r grandissant de gauche à droite

Donc r est le rayon des boules bleues, et un simplexe est considéré dès lors que les boules associées à ces sommets se rencontrent.

Les "trous" dans notre filtration ont un nom plus mathématique que la description que l'on en fait, en effet :

Définition :
Une *classe d'homologie de dimension k (kD)* représente un trou en dimension k qui existe pour une certaine periode d'une filtration.

(Explication mathématique peut etre cool a savoir pour l'oral, mais pas utile dans le projet)

Ainsi une classe d'homologie 0D représente des points connectés, une classe d'homologie 1D représente un trou qui est entouré par un chemin fermé de points connectés (K_2 dans Figure 2 par exemple) et une classe d'homologie 2D représenterait le vide dans une structure de surface fermée.

On note que l'on peut aussi voir les classes d'homologie 0D comme la représentation de "vide" ou "d'absence de lien" entre les composantes connexes. C'est donc pour cela qu'il est important de considérer et les classes d'homologie 0D et les 1D.

Ainsi, grâce à ces définitions, nous sommes capables, depuis un ensemble $X = (x_i)$ de points fini de poids $(w_i)_i$, de créer une filtration et de l'étudier afin de trouver les classes d'homologie 1D qui représentent pour nous les zones critiques de couverture.

I.2 Interprétation de l'homologie persistante

Maintenant que tout cela est plus clair, revenons à notre problème initial. Nous voulons pouvoir détecter les classes d'homologie 1D, c'est à dire les "trous", dans la couverture d'un réseau de transports d'une grande ville. Pour cela, l'homologie persistante nous propose plusieurs affichages graphique afin de rendre compte de ces caractéristiques, nous nous concentrerons sur une seule : *le diagramme de persistance (PD)*

Ce diagramme retrace les "événements" qui sont arrivés lors du parcours d'une filtration. Prenons par exemple le diagramme de persistance associé à la filtration de Figure 2

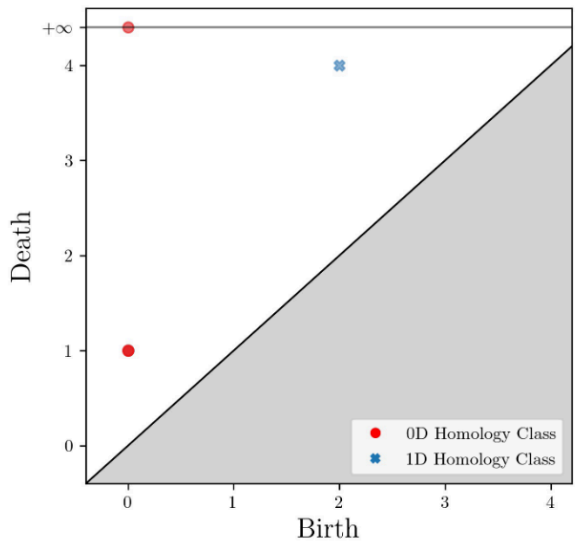


Figure 4: Diagramme de persistance de Figure 2, tiré de [2]

Les chiffres en abscisse et en ordonnée représente l'index du complexe simplicial dans la filtration. En particulier, l'abscisse donne le moment où la classe d'homologie est apparue et l'ordonnée celle où elle disparait (si elle disparait).

Ainsi, nous remarquons qu'en K_0 il y a la naissance de 4 classes 0D (4 composantes connexes) là où en K_1 il n'y en a plus qu'une (d'où la mort de 3 d'entre elle en ordonnée 1, et la dernière qui ne meurt jamais, en $+\infty$). De plus, il y a la création d'un cycle entourant du vide (classe 1D) en K_2 et que celui ci est complétement rempli en K_4.

En reprenant ce qui a été dit précédemment, les classes 0D représentent les composantes connexes vivantes au cours de la filtration, c'est à dire des sous ensembles de stations reliées entre elles. Ainsi un simplexe tuant une homologie 0D (liaison de deux composantes connexes) au rang r représente le fait qu'il est possible à partir de ce rang de se rendre d'une station à l'autre.

Ces simplexes tueurs vont créer des homologies 1D à partir d'un certain moment : des zones entre nos stations reliées. Comprenant qu'il est plus simple de passer par les stations de metros pour aller à une autre plutot que de passer via le centre du cycle.

C'est exactement ce qui nous intéresse : ces zones décrites par les cycles représentent les zones critiques où les personnes sont le moins bien deservies par le réseau de métros, où c'est le plus compliqué de se rendre à une station de métro en prenant en compte le déplacement vers la station (pied ou voiture) et le temps d'attente moyen en station.

II Méthode

De ce qui précède nous pouvons en extraire une méthode générale afin d'analyser un espace métrique pondéré. Celle ci se décompose en 5 étapes :

- Récupération de l'ensemble des points, leurs poids, et leurs distances entre eux
- Creation d'une filtration
- Création de la matrice de bordure
- Réduction de la matrice de bordure
- Construction du diagramme de persistance

La première étape étant développée en Section III, nous supposerons dans la suite de cette section avoir un espace métrique $\left((x_i)_i, d\right)$ tel que chaque x_i admet w_i pour pondération.

À l'étape 2, nous allons créer une filtration grâce à la définition donnée Section I, rien de plus.

Notre but final étant de créer un diagramme de persistance, nous devons reussir à convertir notre filtration en celui ci, cela se fait grâce au théorème centrale :

Définition :
En définissant un espace filtré comme un espace topologique ainsi qu'une de ses filtration, on a :

Théorème : Tout espace vectoriel filtré de dimension finie est isomorphe à la somme directe des espaces filtrés associés à une certaine famille d'intervalles, uniquement définie.

(Intéressant pour comprendre pq on fait tt ça, mais pas compris tous les tenants et aboutissants)

Informatiquement, cela revient à créer une matrice de bordure B, en plaçant un ordre total sur les simplexes du complexe de telle sorte que la face d'un simplexe précède le simplexe et tout simplexe de K_j précède tous les simplexes de K_i tel que $i < j$.

Définition :
On définit la matrice de bordure associée à un ordre total $\sigma_0 < ... < \sigma_n$, en notant n le nombre de simplexes total de la filtration et σ_i un simplexe de la filtration, $\forall (i,j) \in [\![0, n-1]\!]^2$,

$$B[i][j] = \begin{cases} \text{Vrai si } \sigma_i \text{ est une face de } \sigma_j \\ \text{Faux sinon} \end{cases}$$

Un exemple d'une telle matrice est donnée plus bas.

Après avoir calculé B, nous voulons la "réduire" en "code barre", dans le sens où l'on peut interpréter correctement les valeurs de cette matrice avec la filtration (Grâce au théorème énoncé plus tôt). Cet algorithme est nommé *standard algorithm* et est décrit dans [3] par, en posant $\text{low}_B(j) = \max(\{i \in [\![0, n-1]\!], B[i][j] \neq 0\})$:

```
StandardAlgorithm(B) (Réduire une matrice de bordure en code barre)
    for j in [|0, n-1|]:
        while (il existe i < j avec low_B(i) = low_B(j)):
            ajouter colonne i de B à colonne j
```

Comparons alors nos deux matrices, sur l'exemple de la filtration de Figure 2 :

	0	1	2	3	4	5	6	7	8	9
0										
1										
2										
3										
4										
5										
6										
7										
8										
9										

Table 1: Matrice B non réduite

	0	1	2	3	4	5	6	7	8	9
0										
1										
2										
3										
4										
5										
6										
7										
8										
9										

Table 2: Matrice B reduite

Nous pouvons interpréter la matrice $\overline{B}$, si $\text{low}_{\overline{B}}(j) = i$ est défini alors on a une paire de simplexe (σ_i, σ_j) tel que l'apparition de σ_i cause l'apparition d'une classe d'homologie et vient tuer σ_j en apparaissant.

En revanche si $\text{low}_{\overline{B}}(j)$ n'est pas défini alors son apparition cause la naissance d'une classe d'homologie, s'il existe k tel que $\text{low}_{\overline{B}}(k) = j$ on est dans le cas précédent, si k n'existe pas alors la classe d'homologie n'est jamais tuée.

C'est depuis cette matrice $\overline{B}$ réduite que l'on construit notre diagramme de persistance comme il suit :

Définition :
Un diagramme de persistance PD est un multi-ensemble de $\overline{\mathbb{R}^2}$ tel que depuis une matrice réduite $\overline{B}$ on ait, en notant $\mathrm{dg}(\sigma) = l$ si σ apparait à partir de K_l :

$$\mathrm{PD} = \{(\mathrm{dg}(i), \mathrm{dg}(j)), \text{tels que } \mathrm{low}_{\overline{B}}(j) = i\} \cup \{(\mathrm{dg}(i), +\infty), \text{tels que } \mathrm{low}_{\overline{B}}(i) \text{ n'est pas défini}\}$$

III Les données

III.1 Sources

En recherche, le nerd de la guerre c'est les données, ne voulant pas me baser sur des villes factices, j'ai alors décidé de trouver des sources pouvant me fournir des informations sur les stations de metros de plusieurs grandes villes de France comme Paris, Toulouse, Marseille ou même Rennes.

Ainsi toutes les informations relatives aux stations de metros ainsi que les passages sont trouvables via le site du gouvernement : https://transport.data.gouv.fr.

Ces informations servent à définir nos points et notre pondération (voir Section III.2), en revanche elles ne permettent pas d'obtenir les distances entre les stations, pour cela nous utiliserons alors https://www.geoapify.com qui nous renvoie depuis des coordonnées geographiques des temps de trajets en voiture et à pied.

De plus, pour la distance nous avons besoin du nombre d'habitants par arrondissement, pour cela nous utiliserons : implémenté sans cette donnée

III.2 Construction des informations importantes

Définissons dès lors nos objets.

Définition :
Un point x_i, représentant une station de métro, est défini par deux données, celle de la position géographique (latitude/longitude) ainsi que son poids w_i. Le poids w_i est égal à la moyenne du temps d'attente entre deux métros en station x_i sur une semaine entière.

Les temps de passage des metros en station étant plus ou moins constant sur la semaine, il est cohérent d'utiliser une moyenne.

De plus, dans un premier temps, nous définissons similairement à [2] une distance non symétrique entre deux stations x et y :

$$\tilde{d}(x, y) = \min(t_{\text{marche}}(x, y), t_{\text{voiture}}(x, y))$$

Avec $t_{\text{marche}}(x, y)$ le temps qu'il faut en marchant pour aller de la station x à la station y, de même en voiture pour $t_{\text{voiture}}(x, y)$.

On définit finalement la distance (qui cette fois est symétrique):

Définition :
On définit la distance entre deux stations de métros x et y comme :

$$d(x,y) = \frac{1}{P}\Big(P(x)\tilde{d}(x,y) + P(y)\tilde{d}(y,x)\Big)$$

En notant $P(x)$ la population de l'arrondissement de la station x, et $P = P(x) + P(y)$ la somme des population des arrondissement de x et y.

(Temporairement P(x) = 1 $\forall x$, donc $d = \tilde{d}$)

Ainsi en revenant aux boules des complexes simplicaux de Vietoris-Rips, elle relate du coût en temps de prendre le métro. En particulier, $d(x,y)$ est une estimation de la moyenne de temps de trajet d'un individu dans l'arrondissement de la station x allant de x à y et de revenir à x.

Nous pouvons alors analyser les réseaux de transport metropolitain français.

IV Résultats et conclusion

Nous avons choisi de baser notre étude sur les réseaux métropolitain de Paris, Toulouse et Marseille. Leur disposition spatiale dû aux différentes geographies des villes ainsi que leur différence de taille de couverture étant un avantage avoir des résultats pertinents différents.

Ville	Dimension	Médiane	Variance
Paris	0D Homologie	NaN	NaN
	1D Homologie	NaN	NaN
Toulouse	0D Homologie	211.00s	22.27s
	1D Homologie	318.00s	58.62s
Marseille	0D Homologie	184.00s	23.57s
	1D Homologie	223.50s	10.5s

Table 3: Tableau récapitulant les médianes ainsi que la variance des temps de mort des classes holomogiques pour chaque ville.

On comprend que globalement il faut 200s (soit 3m20 environ) pour quelqu'un de se rendre d'une station à une autre (le minimum en temps entre la voiture et la marche) ce qui est effectivement cohérent avec la réalité. En revanche, je n'ai pas encore trouvé d'interprétation aux temps des classes d'homologies 1D.

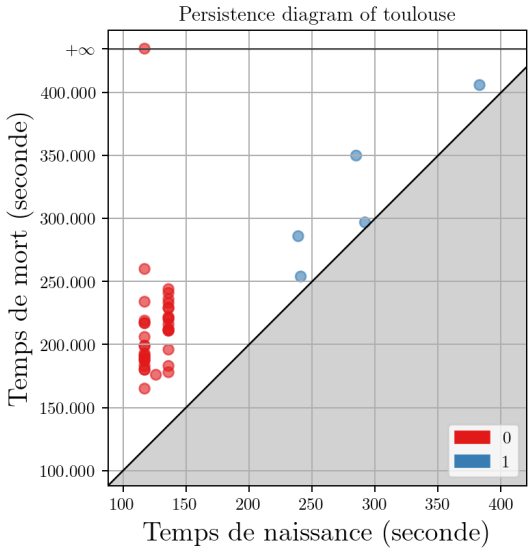


Figure 5: Toulouse

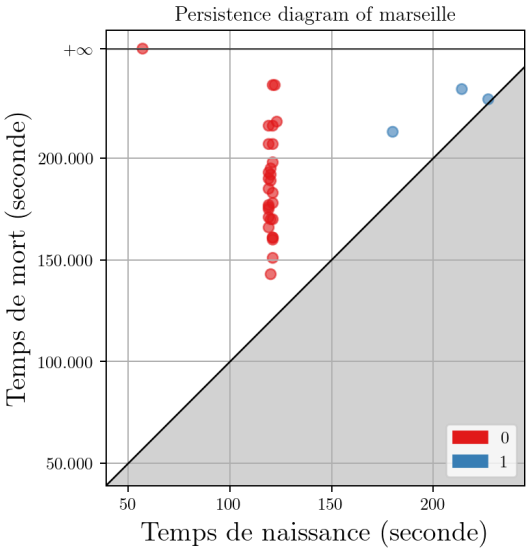


Figure 6: Marseille

Ainsi via ces diagrammes de persistance, on remarque que les stations de metros pour ces deux villes sont égalements réparties en terme de temps de trajet entre deux stations (les classes 0D en rouge). Mais l'interprétation des diagrammes de persistance est assez limité dans notre cas, analysons alors directement les classes 1D se faisant tuer directement sur une carte :

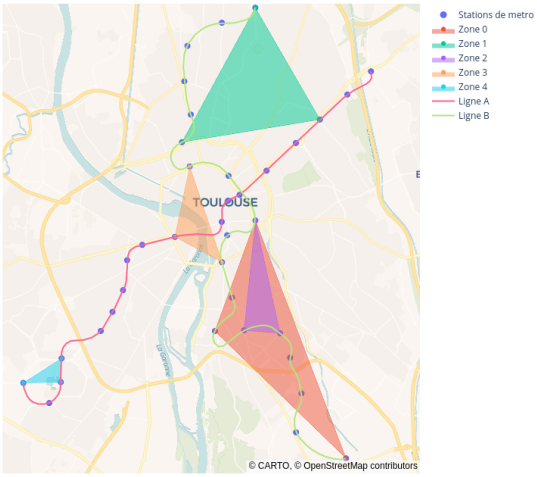


Figure 7: Carte de toulouse

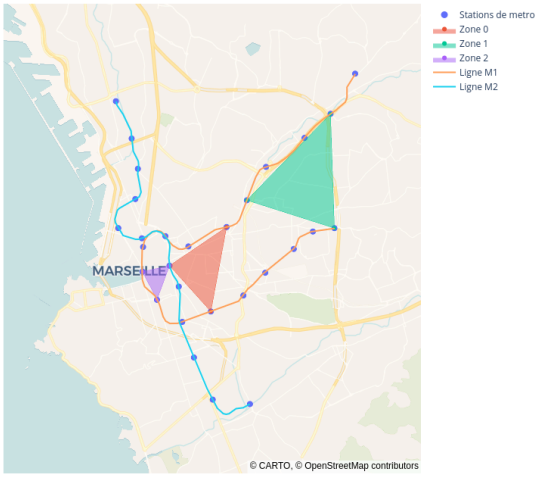


Figure 8: Carte de Marseille

Les triangles ici représentés montrent les zones où il est le plus difficile pour se rendre à une station de métro. Pour les triangles les plus gros, il peut être cohérent de croire qu'il est difficile de se rendre à ces stations de métros en revanche pour les plus petits comme à la Canebière à Marseille cela est plus dur.

Ce sont des zones où il ne circule que très peu de voitures entre les stations de métros, en effet ces zones sont uniquement pietonnes donc la distance parcourue à durée égale est necéssairement plus long à pied qu'en voiture. Donc la distance prise par notre algorithme est

celle relevant de la marche à pied, d'où les zones *a priori* plus petites que celles discutées plus haut.

Nous pouvons observer le plus gros problème de cette méthode : la méthode est pertinente pour le développement d'un réseau autre que le métropolitain (réseau de bus, par exemple). En effet, on remarque que les zones critiques sont entre les lignes de metros dessinées, mais jamais en bout de ligne, là où pourtant la disponibilité des métros est plus faible que dans l'hypercentre des villes.

Ainsi, cette méthode d'analyse peut être pertinente lors d'une simulation pour la création ou l'amélioration prévue d'un réseau, afin de detecter les zones qui seront le plus en besoin avec le réseau imaginé, mais ne permet pas d'établir un tracé *optimal* d'une ligne de métro pour satisfaire le plus de monde.

Bibliography

[1] Henri Paul de Saint-Gervais, *Une Invitation À L'homologie Persistante*, (n.d.)

[2] ABIGAIL HICKOK, BENJAMIN JARMAN, MICHAEL JOHNSON, JIAJIE LUO, and MASON A. PORTER, *Persitent Homology for Resource Coverage: A Case Study of Access to Polling Sites*

[3] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington, *A Roadmap for the Computation of Persistent Homology*