

Ανάλυση Κοινωνικών Δικτύων

2η Εργαστηριακή Άσκηση

Ανίχνευση κοινοτήτων (Community detection)

Συμεών Παπαβασιλείου (papavass@mail.ntua.gr)

Ειρήνη Κοιλανιώτη (eirinikoilanioti@mail.ntua.gr)

Ιωάννης Τζανεττής (gtzane@gmail.com)

Βασίλειος Καρυώτης (vassilis@netmode.ntua.gr)

Κωνσταντίνα Σακκά (nsakka@cn.ntua.gr)

Γεώργιος Καλλίτσης (giorgoskallitsis99@gmail.com)

Επισκόπηση

- Μελέτη χαρακτηριστικών των κόμβων **πραγματικών δικτύων** με χρήση των **μετρικών** που χρησιμοποιήθηκαν στην 1^η άσκηση.
- **Ανίχνευση κοινοτήτων** ως πρόβλημα **διαμέρισης** (partition) ενός γράφου με τους αλγορίθμους
 - Spectral Clustering
 - Newman-Girvan
 - Modularity Maximization
- **Αξιολόγηση διαμερίσεων** με τις μετρικές
 - modularity
 - performance

Πραγματικά δίκτυα

- **American College Football $G_1(V_1, E_1)$**

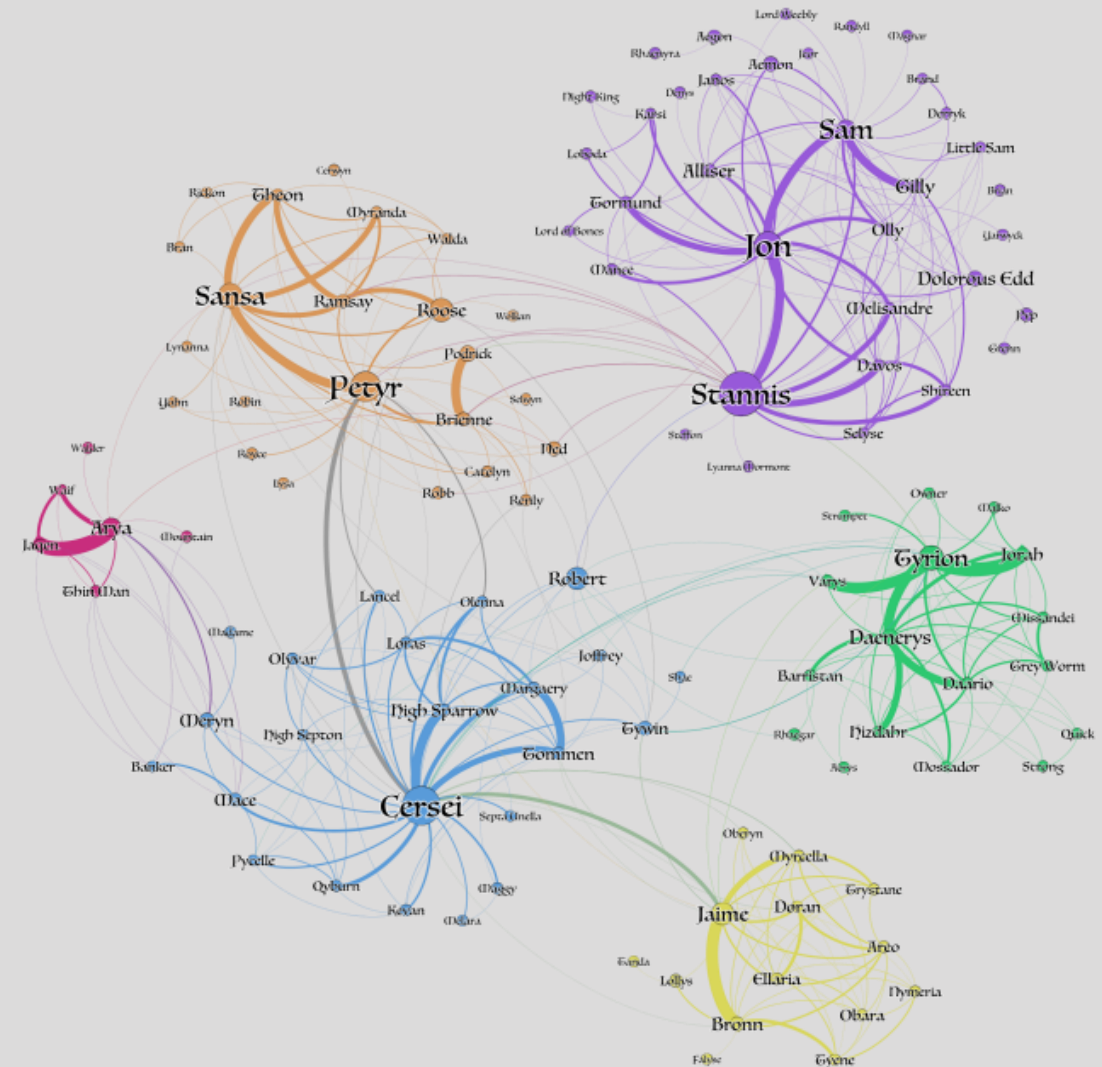
- Αγώνες αμερικάνικου ποδοσφαίρου μεταξύ κολλεγίων κατά το φθινόπωρο του 2000
- $|V_1|=115, |E_1|=613$
- <http://konect.cc/networks/dimacs10-football/>

- **Game of Thrones S5 $G_2(V_2, E_2, w)$**

- $|V_2|=114, |E_2|=396, \sum w_i = 5.139$ αλληλεπιδράσεις μεταξύ των χαρακτήρων της σειράς
- 6 κοινότητες
- <https://networkofthrones.wordpress.com/the-series/season-5/>

- **Email-Eu-core network $G_3(V_3, E_3)$**

- e-mail data από ένα Ευρωπαϊκό ερευνητικό κέντρο
- $(u, v) \in E_3$ αν ο χρήστης u έχει στείλει στον v τουλάχιστον ένα e-mail.
- $|V_3|=1.005, |E_3|=16.706$,
- 42 κοινότητες
- <https://snap.stanford.edu/data/email-Eu-core.html>



Μελέτη χαρακτηριστικών των κόμβων στα πραγματικά δίκτυα

- Βαθμός κόμβου (degree)
- Συντελεστής ομαδοποίησης κόμβου (clustering coefficient)
- Κεντρικότητα εγγύτητας κόμβου (closeness centrality)

Τα αποτελέσματα θα πρέπει να συγκριθούν με τα αντίστοιχα για τις συνθετικές τοπολογίες της προηγούμενης άσκησης.

Ανίχνευση κοινοτήτων (Community detection)

- Η εξέλιξη των σύνθετων δικτύων και οι αλληλεπιδράσεις των συστατικών τους μερών έχουν ως αποτέλεσμα το **σχηματισμό κοινοτήτων**.
- **Κοινότητες** είναι ομάδες κόμβων που παρουσιάζουν κοινές ιδιότητες ή/και έχουν κοινό ρόλο/λειτουργία μέσα στο δίκτυο (π.χ. κόμβοι που αλληλεπιδρούν μεταξύ τους πιο συχνά σε σχέση με άλλους κόμβους, συνιστούν μια κοινότητα)
- Η **ανίχνευση κοινοτήτων** σε έναν γράφο ως **πρόβλημα διαμέρισης** του γράφου:
 - **Βασική ιδέα:** Προσδιορισμός υποσυνόλων κόμβων του γράφου με περισσότερες ακμές μεταξύ των κόμβων κάθε υποσυνόλου παρά εκτός αυτού.
- Η **ποιότητα της διαμέρισης του γράφου** αξιολογείται με διάφορες μετρικές
 - modularity
 - performance

Modularity

- **Βασική ιδέα:** Ένας τυχαίος γράφος δεν αναμένεται να έχει δομή κοινότητας (cluster structure).
- Η μετρική **modularity** εξετάζει το ποσοστό των ακμών που υπάρχουν μέσα σε μια κοινότητα σε σχέση με το ποσοστό των ακμών που θα υπήρχαν εντός της κοινότητας, **αν οι ακμές του γράφου είχαν κατανεμηθεί τυχαία, διατηρώντας την κατανομή του βαθμού κόμβου** (configuration model).
- Έστω δίκτυο $G(V,E)$ με $|E|=m$ (ή $2m$ half-edges), διαμέριση $C=\{C_1, \dots, C_k\}$ σε k κοινότητες, πίνακα γειτνίασης $A=(a_{ij})$ και κόμβους $i, j \in V$, με βαθμό d_i, d_j αντίστοιχα. Με βάση το configuration model, η πιθανότητα να επιλεγεί μια half-edge προσκείμενη στον κόμβο j είναι $p_j=d_j/(2m-1)$.
- Η πιθανότητα να σχηματιστεί μια ακμή μεταξύ των i, j είναι $p_i p_j = d_i d_j / (2m-1)$. Συνεπώς, ο αναμενόμενος αριθμός ακμών μεταξύ των κόμβων i, j είναι $P_{ij}=d_i d_j / 2m$, για μεγάλες τιμές του m .

- **Modularity**
$$Q = \frac{1}{2m} \sum_{l=1}^k \sum_{i \in C_l, j \in C_l} a_{ij} - \frac{d_i d_j}{2m}$$

- $Q \in (-1, 1)$

Performance

- **Βασική ιδέα:** προσδιορισμός του αριθμού των ζευγών κόμβων που έχουν **ερμηνευθεί ορθά** από την μέθοδο της διαμέρισης.
 - αριθμός των ζευγών κόμβων που ανήκουν στην ίδια κοινότητα και συνδέονται με ακμή.
 - αριθμός των ζευγών κόμβων που δεν ανήκουν στην ίδια κοινότητα και δεν συνδέονται με ακμή.
- Έστω δίκτυο $G(V,E)$ με $|V|=n$ και διαμέριση $\mathbf{C}=\{C_1,\dots,C_k\}$.
- **Performance**
$$P(\mathbf{C}) = \frac{|\{(i,j) \in E, C_i = C_j\}| + |\{(i,j) \notin E, C_i \neq C_j\}|}{n(n-1)/2}$$
- $P(\mathbf{C}) \in [0,1]$

Spectral Clustering

- Έστω ο γράφος $G(V,E)$, $|V|=n$, με πίνακα γειτνίασης $A=(a_{ij})$ και βαθμό του κόμβου $i \in V$, d_i .
- Πίνακας βαθμού του G , $D=\text{diag}(d_1, \dots, d_n)$.
- Η μέθοδος αξιοποιεί τις **ιδιοτιμές και τα ιδιοδιανύσματα του Laplacian** πίνακα $L=D-A$.
- Οι ιδιοτιμές και τα ιδιοδιανύσματα του Laplacian πίνακα αποκαλύπτουν **ιδιότητες που αφορούν στην διαμέριση του γράφου** (πλήθος συνδεδεμένων συνιστωσών, πυκνότητα γράφου, minimum cut).
- **Μετασχηματισμός** των κόμβων του γράφου σε **σημεία ενός μετρικού χώρου**.
- Επιλογή k ιδιοδιανυσμάτων του πίνακα L για να προσδιοριστούν οι συντεταγμένες των κόμβων σε έναν μετρικό χώρο διάστασης k .
- Το **πρόβλημα της διαμέρισης του γράφου σε k κοινότητες** ανάγεται σε πρόβλημα **ομαδοποίησης n σημείων του μετρικού χώρου διάστασης k** .
- Χρήση αλγορίθμου ομαδοποίησης (k -means) και ανάθεση σημείων σε κοινότητες.

K-means

1. Είσοδος

k : πλήθος κοινοτήτων

$\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^k$: σύνολο σημείων – κόμβων

2. Αρχικοποίηση

Επιλέγονται τυχαία σημεία του χώρου ως κέντρα των k κοινοτήτων : $\mathbf{w}_1, \dots, \mathbf{w}_k \in \mathbb{R}^k$

3. Ταξινόμηση των σημείων-κόμβων στις κοινότητες

$$u_{ij} = \begin{cases} 1, & j = \arg \min_p \|\mathbf{x}_i - \mathbf{w}_p\|^2 \\ 0, & \text{διαφορετικά} \end{cases}$$

4. Υπολογισμός των νέων κέντρων των κοινοτήτων

$$\mathbf{w}_j := \frac{\sum_{i=1}^n u_{ij} \mathbf{x}_i}{\sum_{i=1}^n u_{ij}}$$

5. Κριτήριο σύγκλισης

Επανάληψη των βημάτων 2 και 3 έως ότου:

- δεν μεταβάλλεται η ταξινόμηση των σημείων-κόμβων ή
- δεν μεταβάλλονται τα κέντρα των κοινοτήτων ή
- ολοκληρώνεται ένας προκαθορισμένος αριθμός επαναλήψεων

Αλγόριθμος Newman - Girvan

- Χρησιμοποιεί την μετρική Edge Betweenness Centrality (EBC).
- Η EBC είναι αντίστοιχη της μετρικής Betweenness Centrality που αφορά τους κόμβους ενός γράφου.
- Η πιο κεντρική ακμή ως προς την EBC: η ακμή που συμμετέχει στο μεγαλύτερο ποσοστό συντομότερων μονοπατιών.
- Ακμές με μεγάλο betweenness centrality είναι συχνά "γέφυρες", δηλαδή, συνδέουν κόμβους που ανήκουν σε διαφορετικές κοινότητες. Η αφαίρεσή τους αποσυνδέει τον γράφο.
- Ο αλγόριθμος Newman-Girvan είναι ένας επαναληπτικός αλγόριθμος για τον εντοπισμό κοινοτήτων.
- Σε κάθε επανάληψη, υπολογίζει την EBC κάθε ακμής (\Rightarrow υψηλό κόστος -χρόνος εκτέλεσης- υπολογισμού της διαμέρισης) και αφαιρεί την ακμή με την μεγαλύτερη EBC.
- Τερματισμός αλγορίθμου: επαναλήψεις έως ότου κάθε κόμβος αποτελεί μία κοινότητα.

Modularity Maximization

- Το πρόβλημα της εύρεσης της διαμέρισης ενός γράφου που **μεγιστοποιεί τη μετρική modularity είναι NP-complete** (Brandes et al., 2006).
- Για την επίλυση του προβλήματος χρησιμοποιούνται ευριστικές μέθοδοι.
- Άπληστη μέθοδος Clauset-Newman-Moore (2004):
 1. Κάθε κόμβος συνιστά μία κοινότητα.
 2. Για κάθε ζεύγος γειτονικών κόμβων, εξετάζεται αν η ανάθεσή τους στην ίδια κοινότητα αυξάνει το modularity. Αν ναι, τοποθετούνται στην ίδια κοινότητα.
 3. Το βήμα 2 επαναλαμβάνεται έως ότου βρεθεί μια τοπικά μέγιστη τιμή modularity.

Ζητούμενα άσκησης

- Να εκτελέσετε τους 3 αλγορίθμους community detection για όλες τις τοπολογίες (πραγματικές και συνθετικές).
- Να οπτικοποιήσετε τις διαμερίσεις.
- Να σχολιάσετε τα αποτελέσματά σας (σύγκριση του αριθμού των κοινοτήτων που υπολογίζει ο κάθε αλγόριθμος, modularity, performance της κάθε διαμέρισης, σύγκριση διαμερίσεων με ground-truth κοινότητες -όπου αυτές δίνονται-, trade-off ποιότητας διαμέρισης και χρόνου εκτέλεσης αλγορίθμων).

Όλοι οι αλγόριθμοι υπάρχουν υλοποιημένοι στην βιβλιοθήκη networkx!