Παραδοσιακές μέθοδοι μάθησης σε γραφήματα

Δημήτριος Κοσμόπουλος

Πανεπιστήμιο Πατρών
Τμήμα Μηχανικών ΗΥ κ Πληροφορικής

8 Δεκεμβρίου 2023

Γραφήματα

Ένα γράφημα

$$\mathcal{G} = (V, \mathcal{E})$$

καθορίζεται από ένα σύνολο κόμβων $\mathcal{V}$ και ένα σύνολο ακμών $\mathcal{E}$ μεταξύ αυτών των κόμβων. Συμβολίζουμε μια ακμή που πηγαίνει από τον κόμβο $v \in \mathcal{V}$ στον κόμβο $u \in \mathcal{V}$ ως

$$(v, u) \in \mathcal{E}.$$

Ταξινόμηση κόμβων

Ταξινόμηση κόμβων σε κοινωνικό δίκτυο: είναι πραγματικός χρήστης ή ρομπότ;

Λάβε υπόψη:

- Ιδιότητες κόμβου
- Γείτονες κόμβου

Διαφορά από κλασσικά προβλήματα επιβλεπόμενης μάθησης: Οι κόμβοι δεν είναι ανεξάρτητοι και ταυτόσημα κατανεμημένοι λόγω της εξάρτησης από τη γειτονιά.

Π.χ. οι άνθρωποι τείνουν να δημιουργούν φιλίες με άλλους που έχουν κοινά ενδιαφέροντα ή δημογραφικά χαρακτηριστικά.

Πρόβλεψη σχέσης (ακμής)

Δίνεται ένα σύνολο κόμβων $\mathcal{V}$ και ένα ατελές σύνολο ακμών μεταξύ αυτών των κόμβων, $\mathcal{E}_{\text{train}} \subset \mathcal{E}$.

Ο στόχος μας είναι να χρησιμοποιήσουμε αυτήν την μερική πληροφορία για να εξάγουμε τις απουσιάζουσες ακμές $\mathcal{E} \setminus \mathcal{E}_{\text{train}}$.

Όπως και η ταξινόμηση κόμβων, η πρόβλεψη σχέσεων διαστρεβλώνει τα όρια των παραδοσιακών κατηγοριών μηχανικής μάθησης, συχνά αναφέρεται τόσο ως εποπτευόμενη όσο και ως μη εποπτευόμενη.

Εντοπισμός κοινότητας

Η ανίχνευση κοινοτήτων είναι η γραφική αντίστοιχη της (μη επιβλεπόμενης) συσταδοποίησης.

Έστω ότι έχουμε πρόσβαση σε όλες τις πληροφορίες παραπομπών στο Google Scholar και δημιουργούμε ένα γράφημα συνεργασίας που συνδέει δύο ερευνητές αν έχουν συντάξει ένα άρθρο μαζί.

Αναμένουμε ότι το γράφημα θα διαχωριστεί σε διάφορες συστάδες (clusters) των κόμβων, που θα είναι ομαδοποιημένες ανά πεδίο έρευνας, θεσμό ή άλλους δημογραφικούς παράγοντες. Με άλλα λόγια, αναμένουμε ότι αυτό το δίκτυο θα εμφανίσει μια δομή κοινοτήτων (community structure), όπου οι κόμβοι είναι πολύ πιθανότερο να συνδεθούν με κόμβους που ανήκουν στην ίδια κοινότητα.

Το πρόβλημα έγκειται στην ανίχνευση τέτοιων κοινοτήτων.

Ταξινόμηση και παλινδρόμηση γραφημάτων

Αντί να προβλέπουμε τα επιμέρους στοιχεία γραφήματος (δηλαδή, τους κόμβους ή τις ακμές), μας δίνεται ένα σύνολο διαφορετικών γραφημάτων και ο στόχος μας είναι να κάνουμε ανεξάρτητες προβλέψεις που είναι ειδικές για κάθε γράφημα. Η ταξινόμηση και η παλινδρόμηση είναι σε αντίστοιχία με την επιβλεπόμενη μάθηση (i.i.d).

- Ταξινόμηση: Δημιουργήστε ένα μοντέλο ταξινόμησης για τον εντοπισμό εάν ένα πρόγραμμα υπολογιστή είναι κακόβουλο, αναλύοντας το γράφημα της συντακτικής του δομής και της ροής δεδομένων.
- Παλινδρόμηση: Δεδομένου ενός γραφήματος που αναπαριστά τη δομή ενός μορίου, δημιουργήστε ένα μοντέλο παλινδρόμησης που θα μπορούσε να προβλέπει την τοξικότητα ή τη διαλυτότητα αυτού του μορίου (πραγματικός αριθμός).

Συσταδοποίηση γραφημάτων

Στο σχετικό έργο της συσταδοποίησης γραφημάτων, ο στόχος είναι να μάθουμε μια μη-εποπτευόμενη μέτρηση ομοιότητας μεταξύ ζευγαριών γραφημάτων.

Με παρόμοιο τρόπο, η ομαδοποίηση γραφημάτων είναι η απλή επέκταση της μη εποπτευόμενης ομαδοποίησης για γραφήματα.

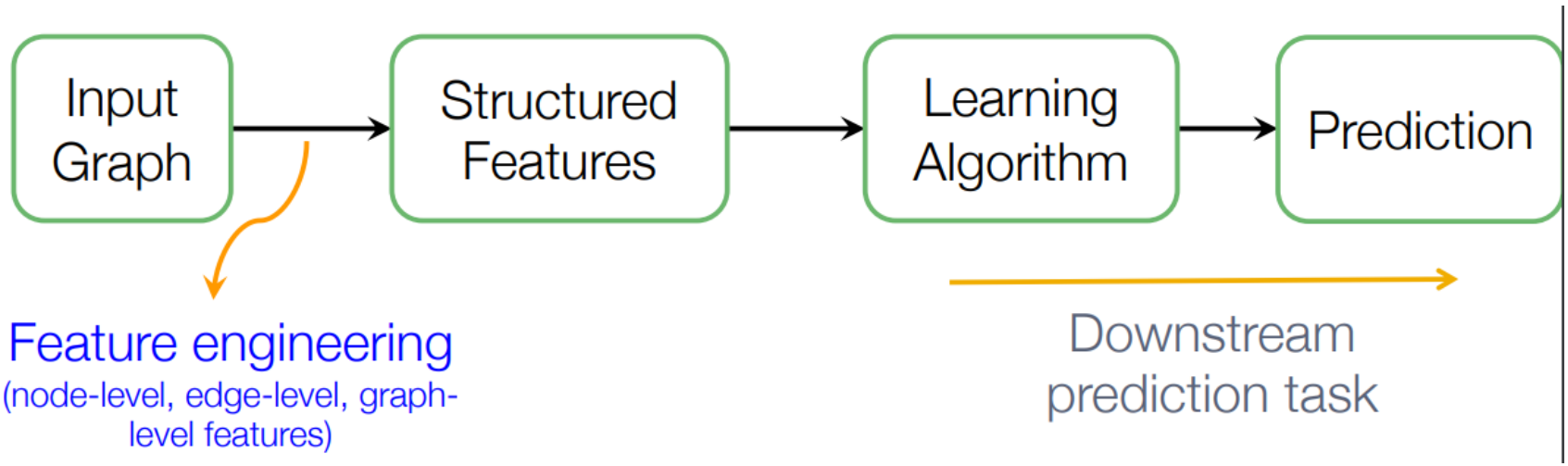
Η πρόκληση σε όλες αυτές τις εργασίες επιπέδου γραφήματος είναι, ωστόσο, πώς να ορίσουμε χρήσιμα χαρακτηριστικά που λαμβάνουν υπόψη τη σχετική δομή μέσα σε κάθε δείγμα δεδομένων.

Δομή

1. Χαρακτηριστικά κόμβων
2. Χαρακτηριστικά ακμών
3. Χαρακτηριστικά γραφημάτων

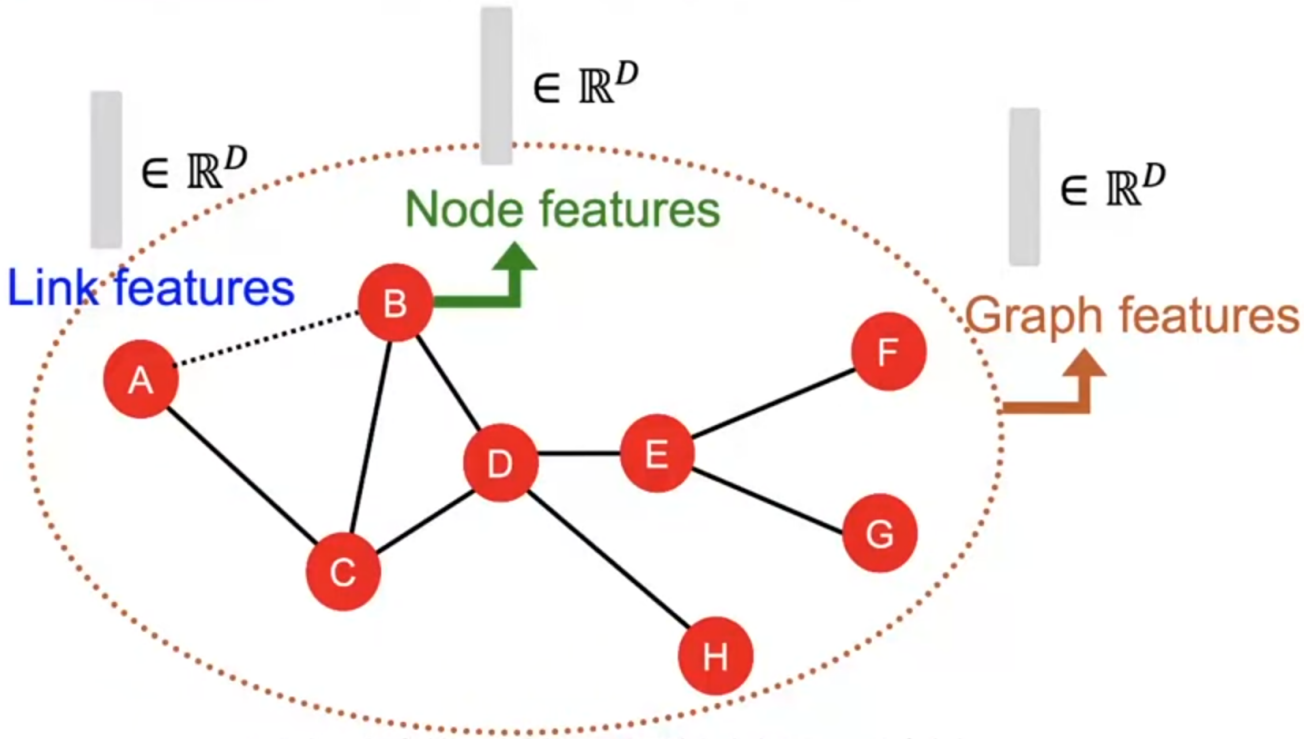
Παραδοσιακές μέθοδοι μηχανικής μάθησης σε γραφήματα

Δεδομένου ενός γραφήματος, εξάγετε χαρακτηριστικά κόμβων, συνδέσμων και επιπέδου γράφου, στη συνέχεια εκπαιδεύετε ένα μοντέλο (SVM, νευρωνικό δίκτυο, κλπ.) που αντιστοιχεί τα χαρακτηριστικά σε ετικέτες.



Παραδοσιακές μέθοδοι μηχανικής μάθησης σε γραφήματα

- Σχεδίαση χαρακτηριστικών για κόμβους/ακμές/γραφήματα
- Υπολογισμός χαρακτηριστικών
- Εκπαίδευση μοντέλου μηχανικής μάθησης (SVM, random forest νευρωνικό, κλπ)
- Δεδομένου διανύσματος χαρακτηριστικών κάνε πρόβλεψη για κόμβους/ακμές/γραφήματα



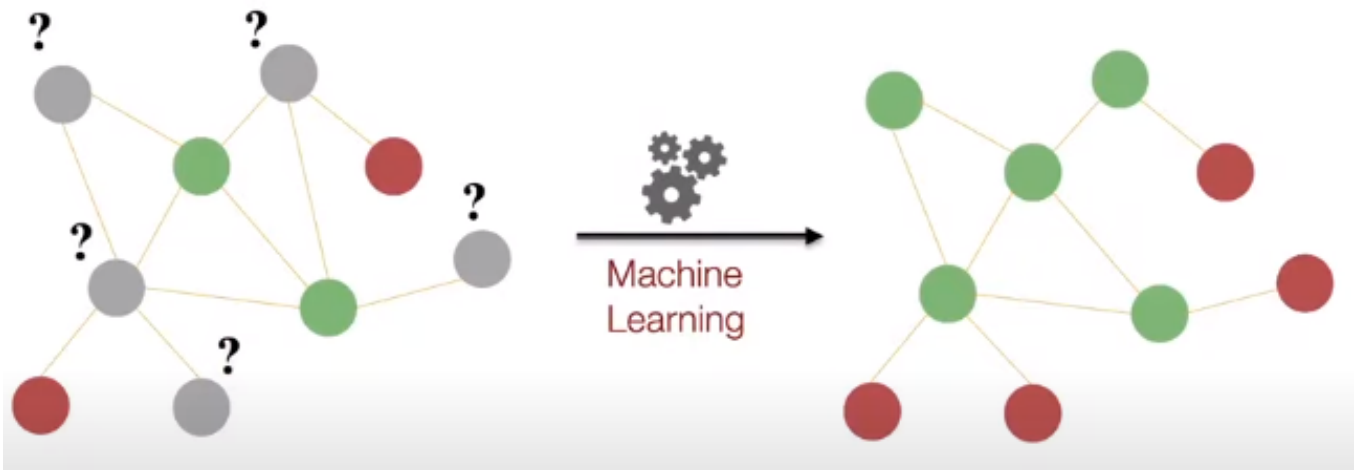
Παραδοσιακές μέθοδοι μηχανικής μάθησης σε γραφήματα

Δεδομένου:

$$\mathcal{G} = (V, \mathcal{E})$$

π.χ. για επίπεδο κόμβων μάθε συνάρτηση:

$$f : V \to \mathbb{R}^d$$



Χαρακτηριστικά κόμβων

- Βαθμός
- Κεντρικότητα
- Συντελεστής συσταδοποίησης
- Graphlets

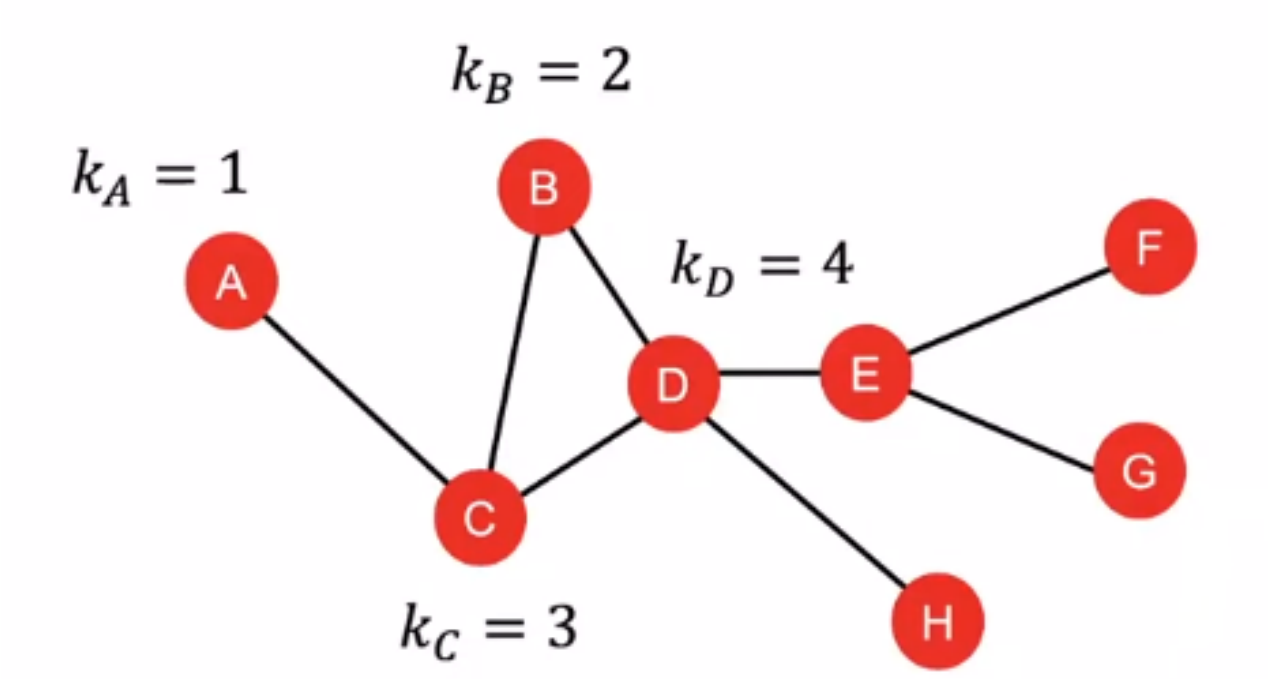
Βαθμός κόμβων

Βαθμός (δεγρεε) κόμβου:

$$k_u = \sum_{v \in V} \mathrm{A}[u, v]$$

Μετράει τον αριθμό ακμών που προσπίπτουν στον κόμβο.

Δεν λαμβάνει υπόψη τη σημαντικότητα των γειτόνων.



Κεντρικότητα ιδιοδιανύσματος - eigenvector centrality

Κεντρικότητα κόμβου:

$$c_u = \frac{1}{\lambda} \sum_{v \in N(u)} c_v \tag{1}$$

Ξαναγράφοντας αυτήν την εξίσωση σε διανυσματική σημειογραφία με το c ως το διάνυσμα των κεντρικοτήτων των κόμβων, μπορούμε να δούμε ότι αυτή η επανάληψη καθορίζει την τυπική εξίσωση ιδιοδιανύσματος για τον πίνακα γειτνίασης A:

$$\lambda \mathrm{c} = \mathrm{Ac} \tag{2}$$

Η μεγαλύτερη ιδιοτιμή είναι πάντα θετική και μοναδική (Perron - Frobenius) και εκφράζει την κεντρικότητα του κόμβου.

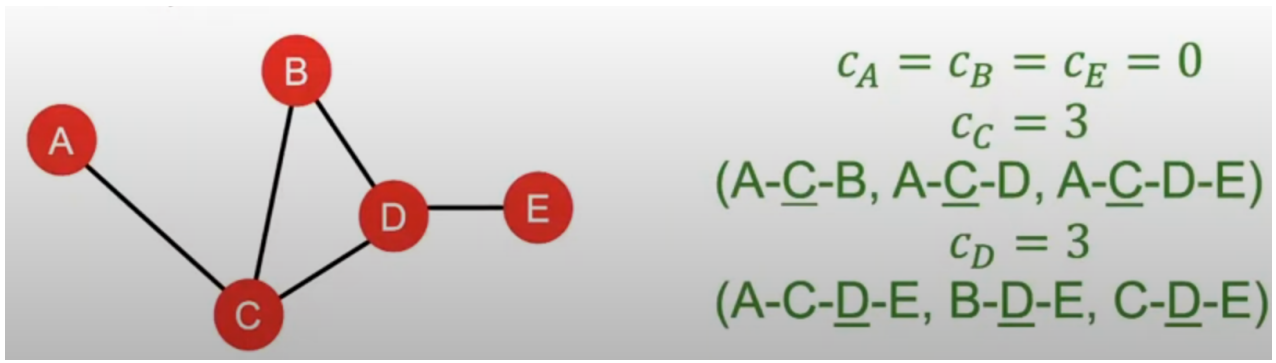
Η αντίστοιχη τιμή του ιδιοδιανύσματος που αντιστοιχεί στη μέγιστη ιδιοτιμή δίνει την κεντρικότητα κάθε κόμβου.

Λαμβάνει υπόψη τη σημαντικότητα των γειτόνων.

Κεντρικότητα μεσολάβησης - betweenness centrality

Ο κόμβος είναι σημαντικός αν βρίσκεται στο συντομότερο μονοπάτι μεταξύ άλλων κόμβων.

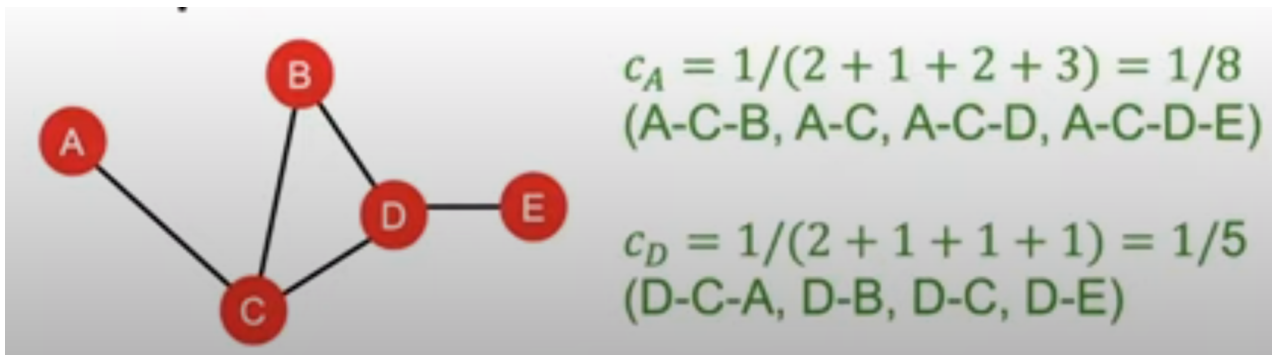
$$c_v = \sum_{s \neq v \neq t} \frac{\#(\textit{shortest paths between } s \textit{ and } t \textit{ that contain } v)}{\#(\textit{shortest paths between } s \textit{ and } t)} \tag{3}$$



Κεντρικότητα εγγύτητας - closeness centrality

Ο κόμβος είναι σημαντικός αν το συντομότερο μονοπάτι προς τους άλλους κόμβους είναι μικρό.

$$c_v = \frac{1}{\sum_{u \neq v} \textit{shortest path length between } u \textit{ and } v} \tag{4}$$



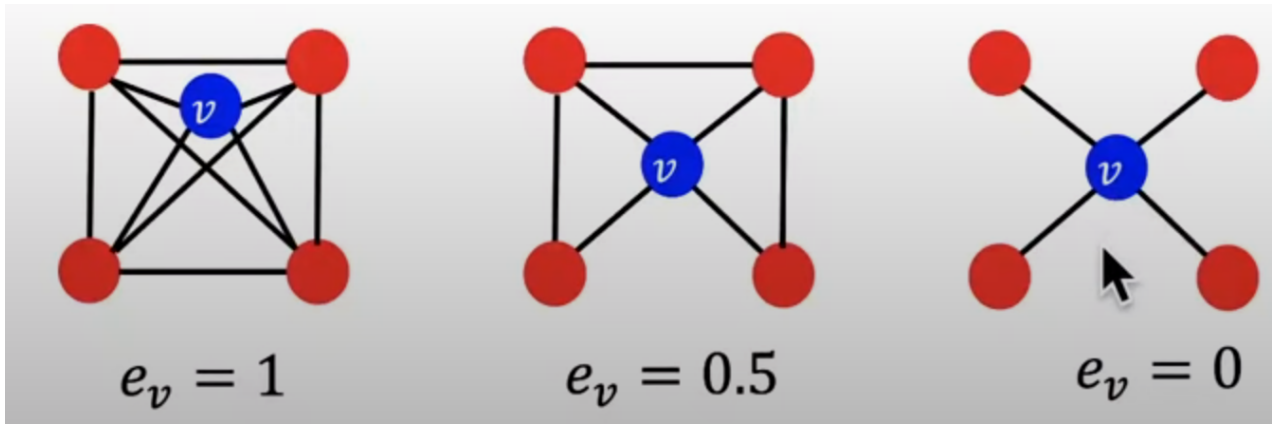
Συντελεστής συσταδοποίησης - clustering coefficient

Μετράει την συνδεσιμότητα των γειτόνων του κόμβου.

$$e_v = \frac{|(u_1, u_2) \in \mathcal{E} : u_1, u_2 \in \mathcal{N}(v)|}{\binom{k_v}{2}} \tag{5}$$

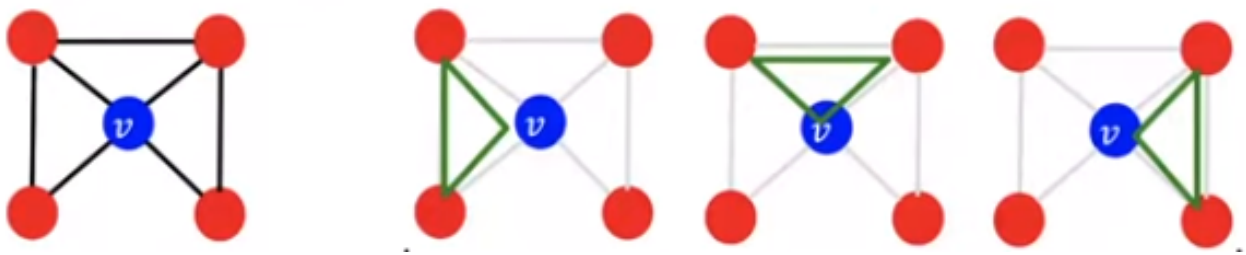
Ο αριθμητής εκφράζει τον αριθμό των ακμών μεταξύ γειτονικών κόμβων.

Ο παρονομαστής εκφράζει τον μέγιστο αριθμό των ακμών μεταξύ γειτονικών κόμβων (συνδυασμοί ανά 2).



Graphlets

Ο συντελεστής συσταδοποίησης πρακτικά μετρά τον αριθμό τριγώνων στο δίκτυο της γειτονιάς.

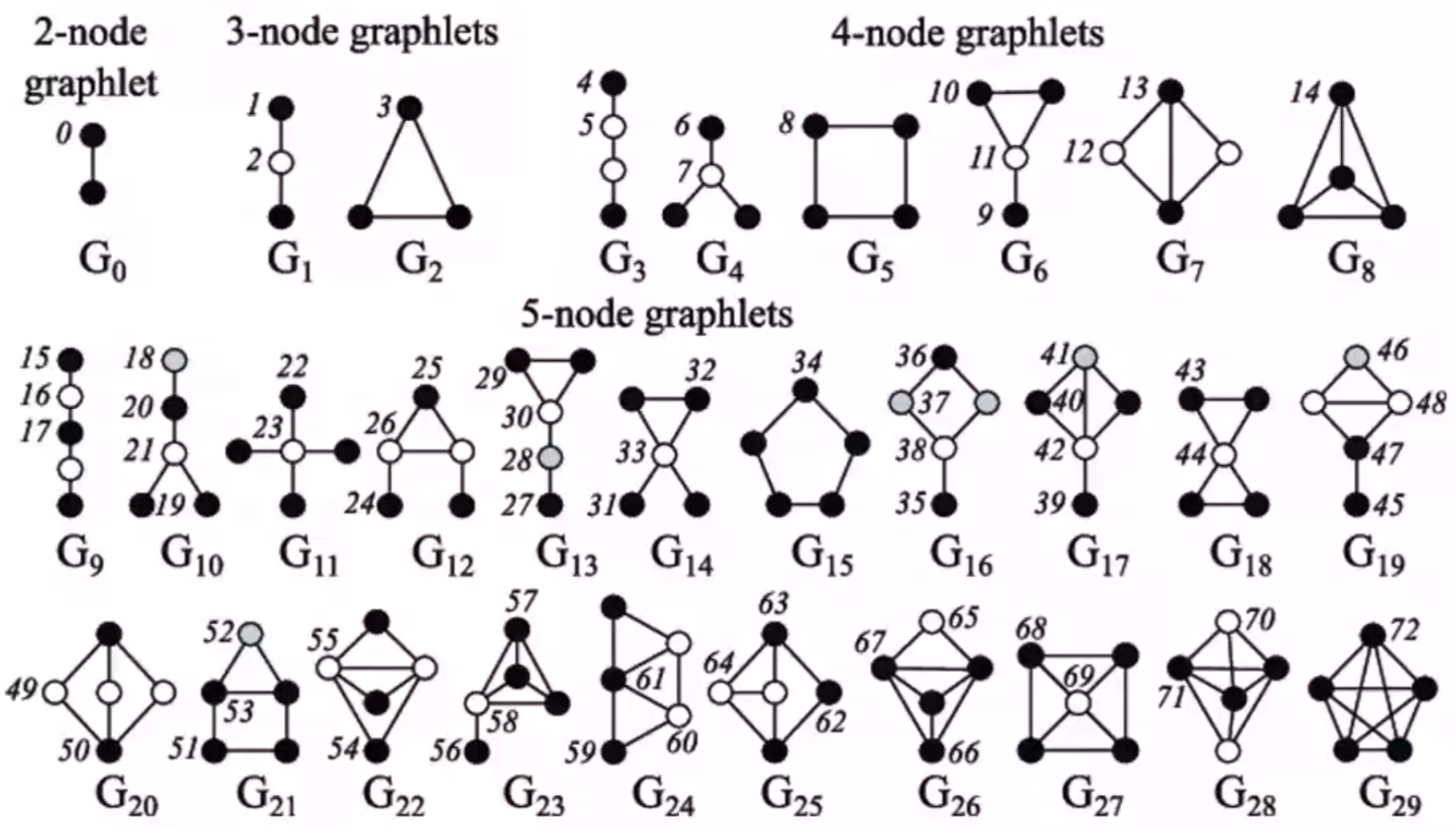


$e_v = 0.5$ διότι υπάρχουν 3 τρίγωνα σε σύνολο 6 (από πλήρως συνδεδεμένο γράφημα).

Μπορούμε να γενικεύσουμε την ιδέα και για άλλα υπο-γραφήματα πλην των τριγώνων (graphlets).

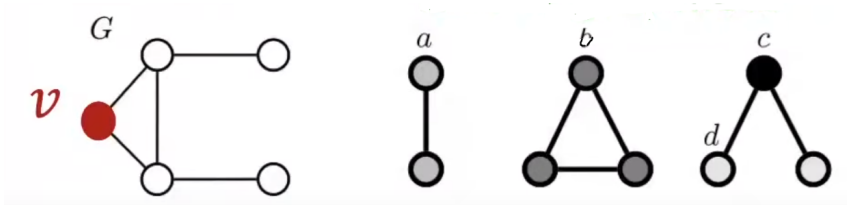
Graphlets

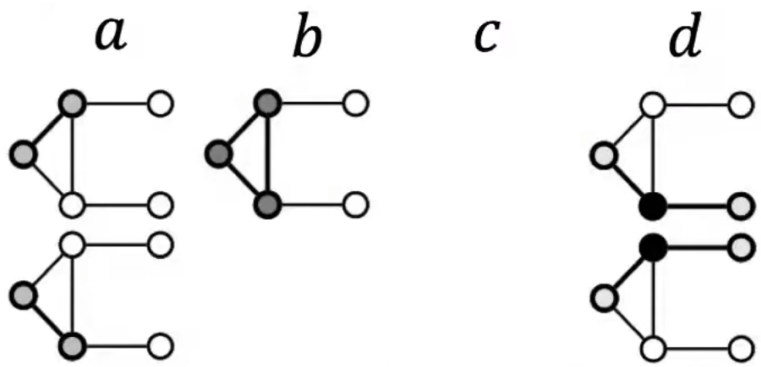
Συνδεδεμένα μη ισομορφικά υπο-γραφήματα με σημείο εκκίνησης (ρίζα).



Graphlets

Διάνυσμα βαθμών graphlet (Graphlet Degree Vector - GDV) : προκύπτει αν θεωρήσουμε τα graphlet ως βάση διανυσματικού χώρου.
Το GDV μετρά τον αριθμό των graphlet στα οποία συμμετέχει ο κόμβος.





Άρα ο κόμβος ν αναπαρίσταται βάσει των [a b c d] ως [2, 1, 0, 2]

Graphlets

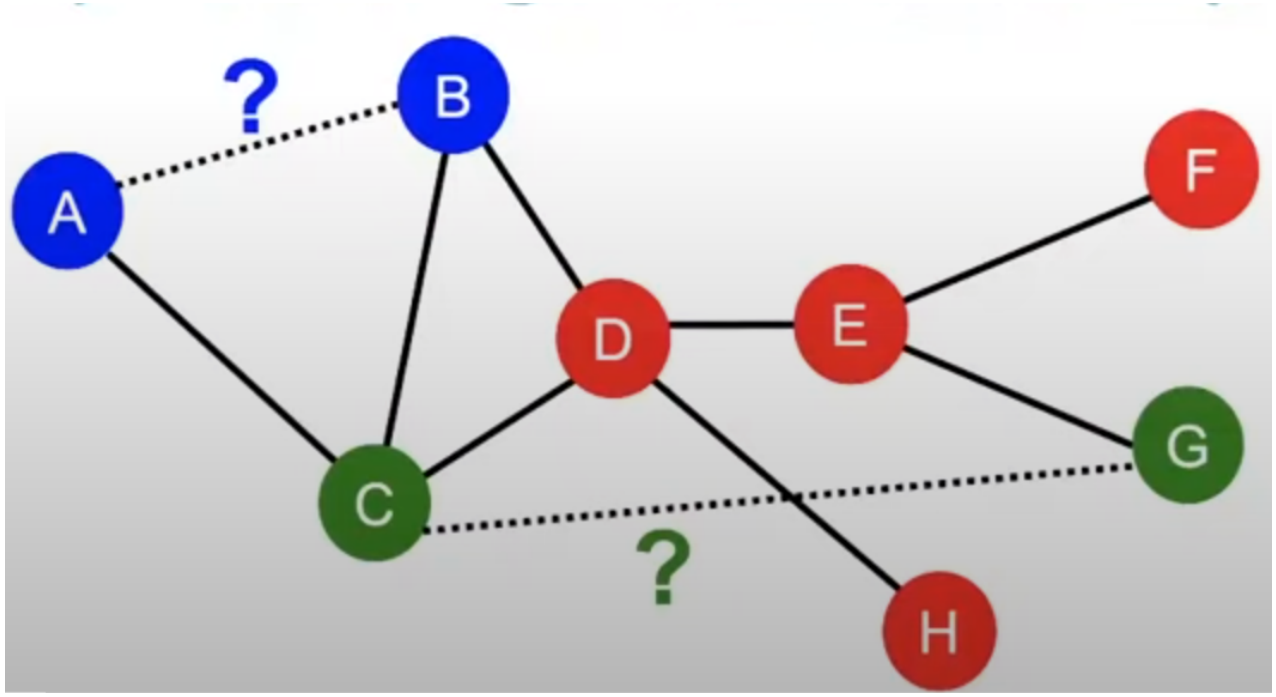
- Χρησιμοποιώντας graphlets μέχρι 5 κόμβους έχουμε διανυσματική αναπαράσταση με 73 παραμέτρους που περιγράφει την τοπολογία της γειτονιάς.
- Περιγράφει την τοπολογία μέχρι 4 βήματα από τον κόμβο.
- Πιο λεπτομερής αναπαράσταση σε σύγκριση με το βαθμό ενός κόμβου.

Σύνοψη χαρακτηριστικών κόμβων

- Χαρακτηριστικά βασισμένα στη σημασία ενός κόμβου. Εύρεση των πιο επιδραστικών κόμβων.
  - Βαθμός κόμβου
  - Κεντρικότητα (ιδιοδιάνυσμα, μεσολάβηση, εγγύτητα)
- Χαρακτηριστικά βασισμένα στη δομή. Εύρεση του ρόλου του κόμβου στη δομή.
  - Βαθμός κόμβου
  - Συντελεστής συσταδοποίησης
  - Graphlet Degree Vector

Πρόβλεψη ακμών

- Πρόβλεψη ακμών με βάση τις ήδη υπάρχουσες
- Όλα τα ζεύγη κόμβων ταξινομούνται και τα πρώτα K αντιστοιχίζονται σε ακμή
- Τα χαρακτηριστικά επομένως θα πρέπει να αφορούν ζεύγη κόμβων



Τυποποίηση προβλήματος

Εκπαίδευση:

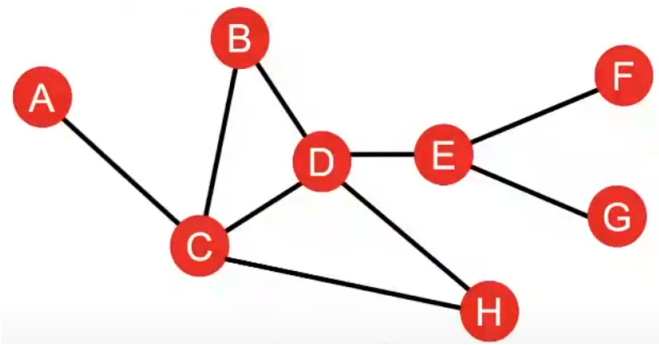
- Για στατικά γραφήματα:
  - Αφαίρεσε αριθμό από ακμές.
  - Προσπάθησε να προβλέψεις τις αφαιρεθείσες ακμές
- Για γραφήματα δυναμικά εξελισσόμενα στο χρόνο:
  - Κάνε πρόβλεψη ταξινομημένης λίστας K ακμών που αναμένεις να εμφανιστουν σε χρονικό διάστημα T
  - Επαλήθευσε την εμφάνιση των ακμών στο χρονικό διάστημα T

Μεθοδολογία

- Για κάθε ζεύγος κόμβων (u, v) υπολόγισε το σκορ $c(u, v)$ π.χ. αριθμό κοινών γειτόνων.
- Ταξινόμησε τα ζεύγη ανάλογα με το σκορ
- Κάνε πρόβλεψη για τα k κορυφαία στη λίστα
- Σύγκρινε με το ποια τελικά εμφανίζονται

Χαρακτηριστικά ακμής βάσει απόστασης

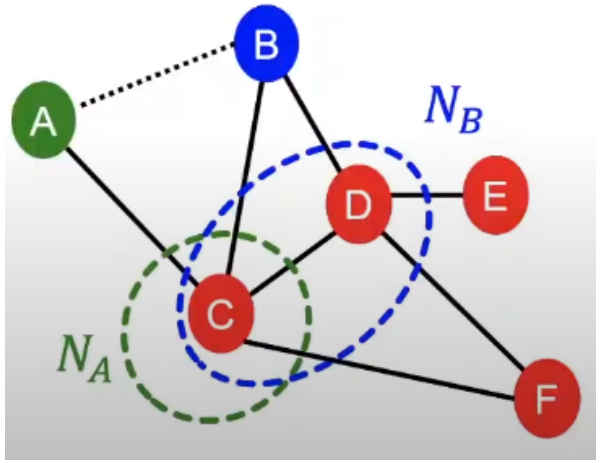
Συντομότερο μονοπάτι μεταξύ 2 κόμβων



$$S_{BH} = S_{BE} = S_{AB} = 2$$
$$S_{BG} = S_{BF} = 3$$

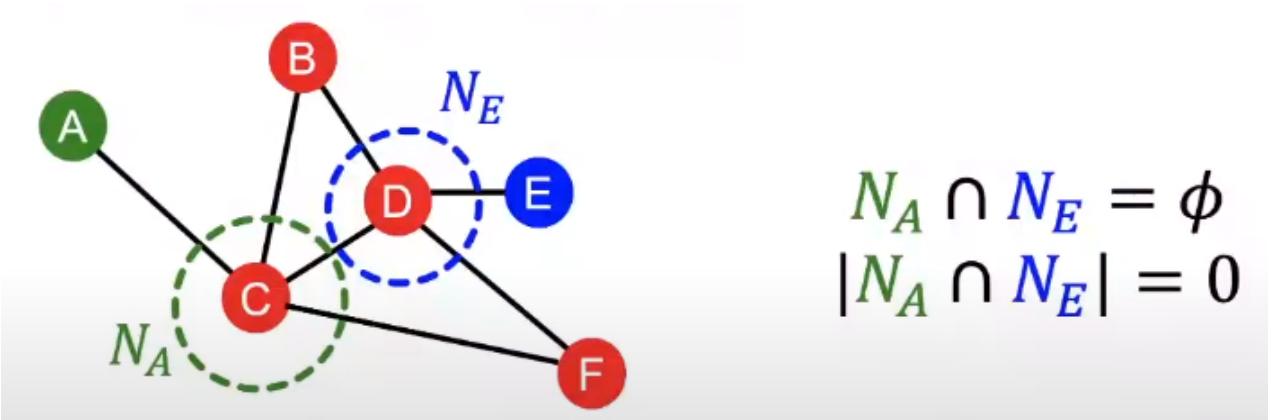
Δεν αναπαριστά τον αριθμό από συντομότερα μονοπάτια π.χ. (B,H) έχει 2 τέτοια μονοπάτια ενώ (B,E) και (A,B) έχουν μόνο 1 (η σύνδεση μεταξύ B,H φαίνεται ισχυρότερη).

Χαρακτηριστικά ακμής βάσει τοπικής γειτονίας



- Κοινοί γείτονες: $|N(v_1) \cap N(v_2)|$
  π.χ. $|N(A) \cap N(B)| = |\{C\}| = 1$
- Συντελεστής Jaccard: $\frac{|N(v_1) \cap N(v_2)|}{|N(v_1) \cup N(v_2)|}$
  π.χ. $\frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|} = \frac{|\{C\}|}{|\{C,D\}|} = \frac{1}{2}$
- Συντελεστής Adamic Adar: $\sum_{u \in N(v_1) \cap N(v_2) \subseteq \bigcup} \log(k_u)$

Χαρακτηριστικά ακμής βάσει συνολικού γραφήματος



- *Περιορισμός:* Οι μετρικές που βασίζονται σε τοπική γειτονία μηδενίζονται εάν δύο κόμβοι δεν έχουν κοινούς γείτονες
- Ωστόσο οι κόμβοι μπορούν να συνδεθούν στο μέλλον
- Εξαγωγή χαρακτηριστικών από όλο το γράφημα θα έλυνε το πρόβλημα

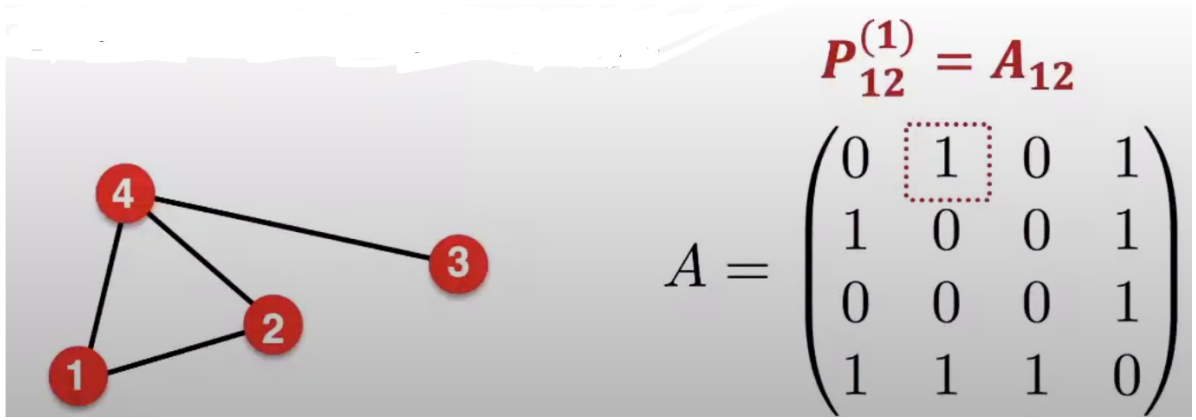
Χαρακτηριστικά ακμής βάσει συνολικού γραφήματος

Δείκτης KATZ: υπολογίζει τον αριθμό των μονοπατιών οποιουδήποτε μήκους μεταξύ δύο ζευγών κόμβων.

- Πώς μετράμε τον αριθμό των μονοπατιών;
- Χρήση δυνάμεων του πίνακα γειτνίασης

Δυνάμεις του πίνακα γειτνίασης

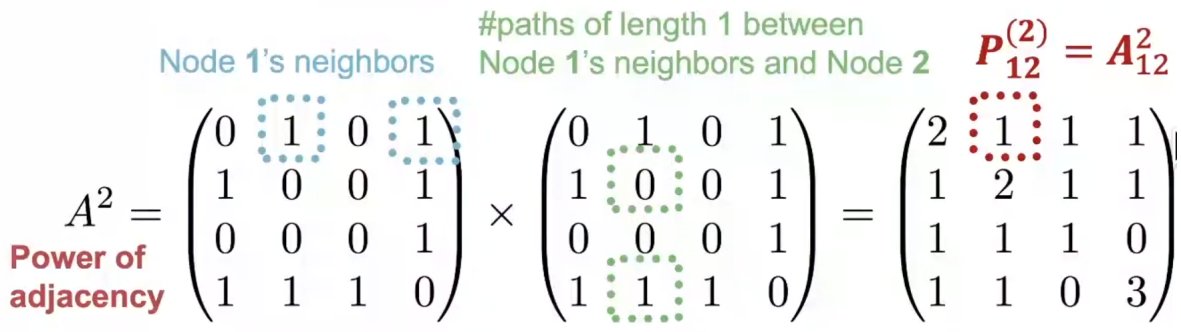
- $A_{u,v} = 1$ αν $u \in N(v)$
- Θα δείξουμε για τον πίνακα που περιέχει τον αριθμό μονοπατιών μήκους k ότι $\mathbf{P}^{(k)} = \mathbf{A}^k$
- $\mathbf{P}^{(1)}_{uv}$ δίνει τον αριθμό των μονοπατιών μεταξύ u,v μήκους 1.



Δυνάμεις του πίνακα γειτνίασης

Πώς υπολογίζουμε $\mathbf{P}_{uv}^{(2)}$;

- Υπολόγισε τον αριθμό μονοπατιών μήκους 1 μεταξύ κάθε γείτονα του u και του v
- Υπολόγισε το άθροισμα για όλους τους γείτονες του u
  $\mathbf{P}_{uv}^{(2)} = \Sigma_{i \in N(u)} A_{ui} * P_{iv}^{(1)} = \Sigma_{i \in N(u)} A_{ui} * A_{iv} = \mathbf{A}_{uv}^2$



- Ομοίως $\mathbf{P}_{uv}^{(l)} = \mathbf{A}_{uv}^{(l)}$ περιγράφει μονοπάτια μήκους l

Δείκτης Katz

Ο δείκτης Katz από το u στο v υπολογίζεται ως:

$$\mathrm{S}_{u,v} = \sum_{l=1}^{\infty} \beta^l \, \mathbf{A}^l_{u,v}$$

όπου $\mathbf{A}^l_{u,v}$ περιγράφει τα μονοπάτια μήκους l από το u στο v και το $0 < \beta < 1$ βάζει ποινή σε μονοπάτια πολύ μεγάλου μήκους.

Ο δείκτης υπολογίζεται σε κλειστή μορφή:

$$\mathbf{S} = \sum_{l=1}^{\infty} \beta^l \mathbf{A}^l = (\mathbf{I} - \beta \mathbf{A})^{-1} - \mathbf{I}$$

Σύνοψη για χαρακτηριστικά ακμών

- Βασισμένα στην απόσταση: χρησιμοποιούν συντομότερο μονοπάτι αλλά δε λαμβάνουν υπόψη την πολλαπλότητα του μονοπατιού
- Τοπική γειτονία: αποτυπώνει την πολλαπλότητα των κοινών γειτόνων αλλά δίνει μηδέν αν δεν υπάρχουν κοινοί γείτονες
- Συνολικό γράφημα: χρησιμοποιεί όλο το γράφημα και το δείκτη Katz για να υπολογίσει όλα τα μονοπάτια μεταξύ δύο κόμβων.

Χαρακτηριστικά σε επίπεδο γραφημάτων

- Πώς αναπαριστούμε γραφήματα π.χ. για ένα πρόβλημα ταξινόμησης;
- Χρήση πυρήνων (kernels) αντί διανυσμάτων χαρακτηριστικών.

Μέθοδοι πυρήνων

- Πυρήνας $K(G, G')$ μετρά ομοιότητα μεταξύ δεδομένων.
- Πίνακας $\mathbf{K} = [K(G, G')]_{G,G'}$ πάντοτε θετικά ημιορισμένος (θετικές ιδιοτιμές).
- Υπάρχει αναπαράσταση χαρακτηριστικών $\phi(\cdot)$ τέτοια ώστε $K(G, G') = \phi(G)^T \phi(G')$.
- Από τη στιγμή που οριστεί ο πυρήνας υπάρχουν έτοιμα μοντέλα π.χ. Kernel-SVM τα οποία μπορούν να χρησιμοποιηθούν.
- ΠΛΕΟΝΕΚΤΗΜΑ: Το πρόβλημα μπορεί να οριστεί βάσει των γινομένων που ορίζει ο πυρήνας και δεν χρειάζεται καν να οριστεί η διανυσματική αναπαράσταση, η οποία μπορεί να έχει και άπειρη διάσταση.

Ταξινόμηση με πυρήνες

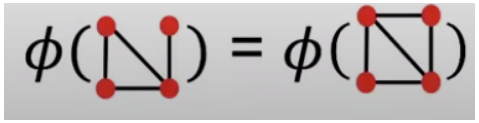
```python
from sklearn.svm import SVC
import numpy as np

# Define your custom kernel function
def custom_kernel(X, Y):
    # Implement your kernel logic here
    # For example, a simple linear kernel could be:
    return np.dot(X, Y.T)

# Create an SVM instance with the custom kernel
clf = SVC(kernel=custom_kernel)

# Train the model
clf.fit(X_train, y_train)

# Make predictions
predictions = clf.predict(X_test)
```

Επιλογές πυρήνων για γραφήματα

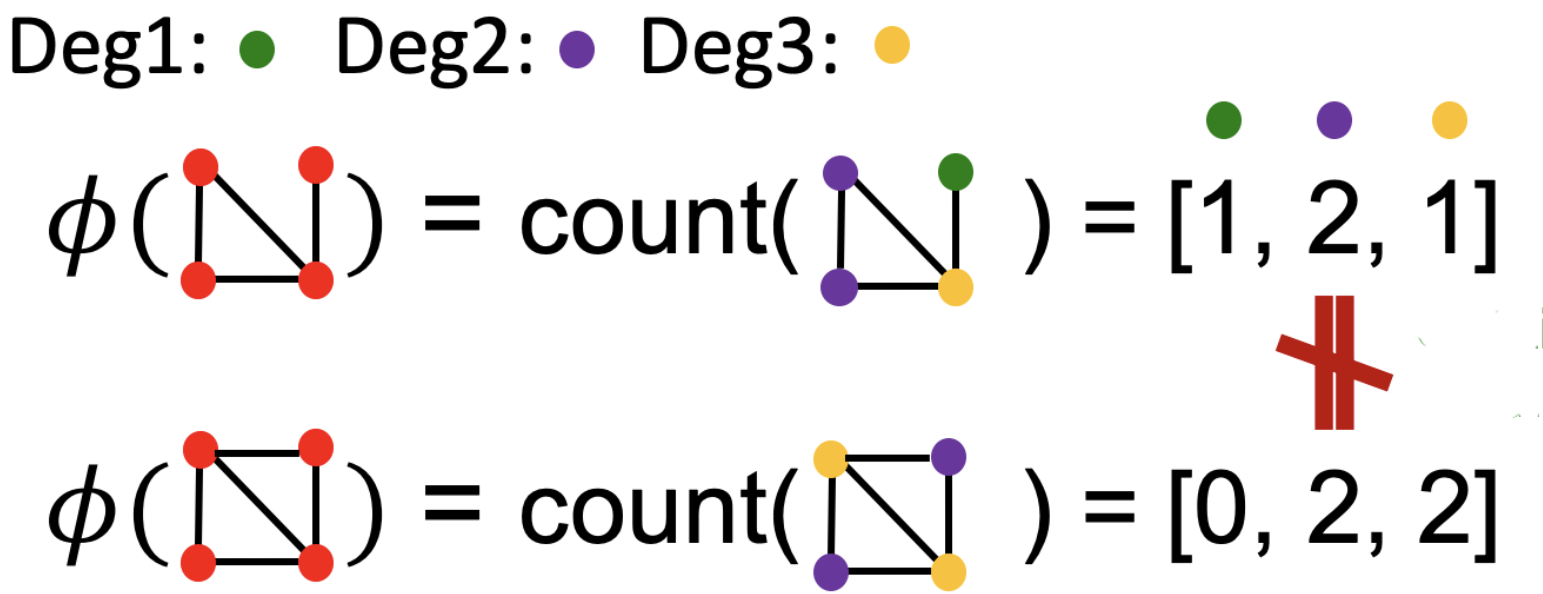
- Graphlet kernel
- Weisfeilel - Lehman kernel
- Random walk kernel
- Shortest path Kernel

Βασική ιδέα πυρήνων για γραφήματα

- bag of words: η διανυσματική αναπαράσταση ϕ είναι το ιστόγραμμα ¨λέξεων' από ένα λεξικό
- Η δομή χάνεται
- Απλοϊκή εφαρμογή: οι κόμβοι είναι οι λέξεις (διαφορετικά γραφήματα $\rightarrow$ ίδιες αναπαραστάσεις)



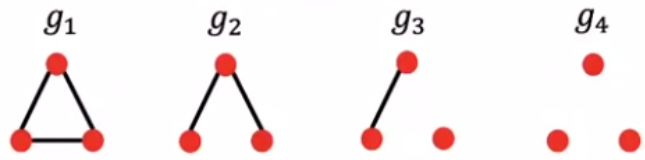
- Οι λέξεις είναι ο βαθμός των κόμβων:



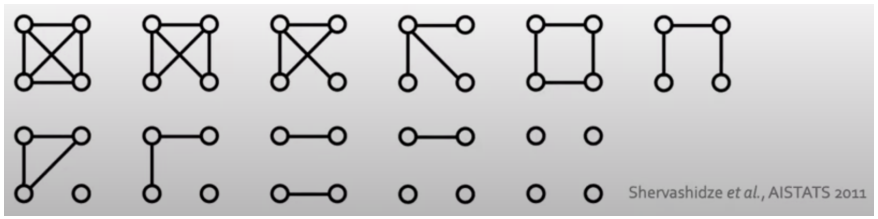
Graphlet kernel

- Μέτρα τον αριθμό Graphlet στο γράφημα
- Ο ορισμός διαφέρει από τα χαρακτηριστικά κόμβων διότι αναφέρεται σε όλο το γράφημα
  - Δε χρειάζεται οι κόμβοι να είναι συνδεδεμένοι
  - Δεν υπάρχει ρίζα

    π.χ. $k=3$

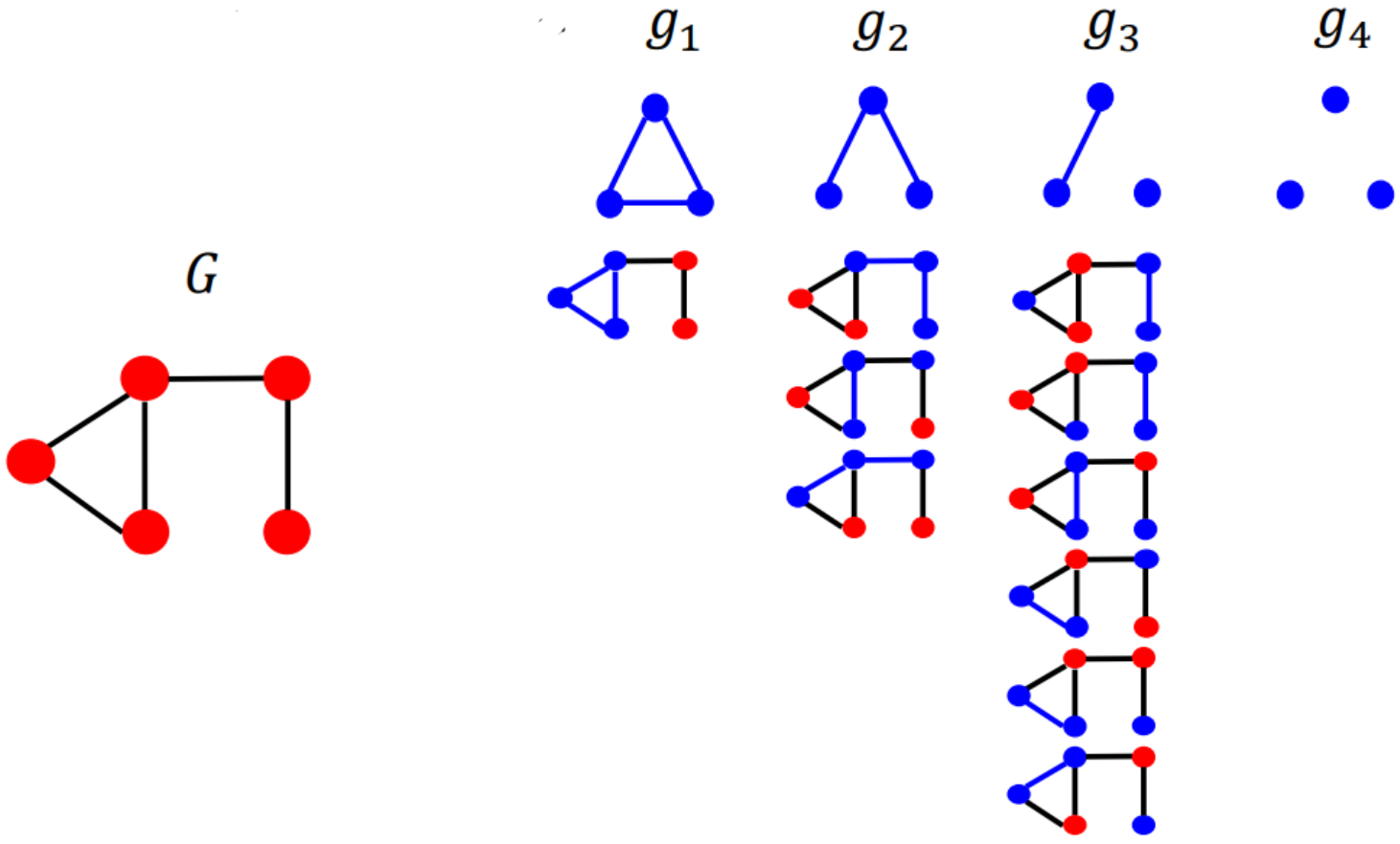


$k=4$



Graphlet kernel

Παράδειγμα για $k=3$



$$f_G = (1, \quad 3, \quad 6, \quad 0)^{\mathrm{T}}$$

Graphlet kernel

Ο πυρήνας υπολογίζεται ως $K(G, G') = {f_G}^{\mathrm{T}} f_{G'}$
ή καλύτερα : $K(G, G') = {h_G}^{\mathrm{T}} h_{G'}$
όπου $h_G = \frac{f_G}{\mathrm{Sum}(f_G)}$ για κανονικοποίηση.

Προβλήματα:

- ο υπολογισμός graphlet μεγέθους k σε γράφημα μεγέθους n χρειάζεται $O(n^k)$ (συνέπεια του NP-hard προβλήματος ισομορφισμού)
- αν ο βαθμός των κόμβων είναι φραγμένος από το d τότε $O(nd^{k-1})$

Weisfeiler - Lehman kernel

Ανάγκη: αλγόριθμος χαμηλού κόστους.

Ιδέα: Χρησιμοποίηση της δομής της γειτονιάς για να φτιάξουμε ένα λεξιλόγιο γενικεύοντας την έννοια του βαθμού ενός κόμβου.

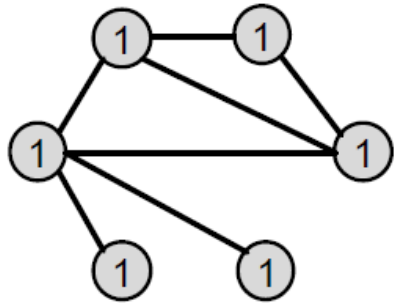
- Με δεδομένο γράφημα $\mathcal{G} = (V, \mathcal{E})$
  - Ανάθεσε χρώμα $c^{(0)}(v)$ σε κάθε κόμβο v
  - Επαναληπτικά ανανέωσε τα χρώματα των κόμβων σύμφωνα με : $c^{(k+1)}(v) = \mathrm{HASH}\left(\left\{c^{(k)}(v), \left\{c^{(k)}(u)\right\}_{u \in N(v)}\right\}\right)$ όπου η συνάρτηση HASH απεικονίζει διαφορετικές εισόδους σε διαφορετικά χρώματα.
  - Έπειτα από k επαναλήψεις $c^{(k+1)}(v)$ αναπαριστά τη γειτονιά έως k βήματα.

Weisfeiler - Lehman kernel

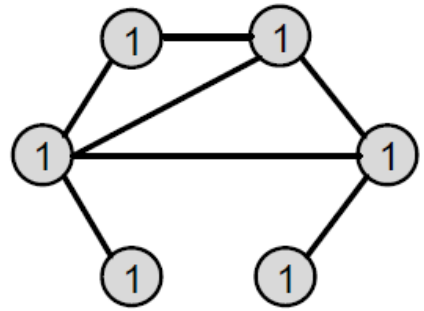
Για 2 γραφήματα:

1. Ανάθεσε αρχικά χρώματα:

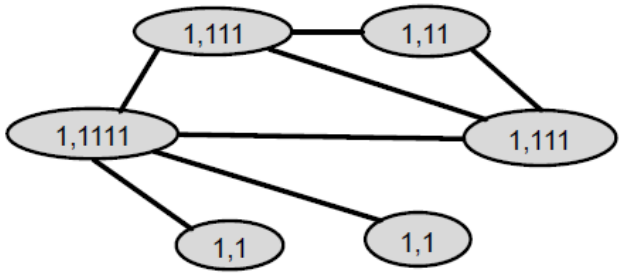
G_1

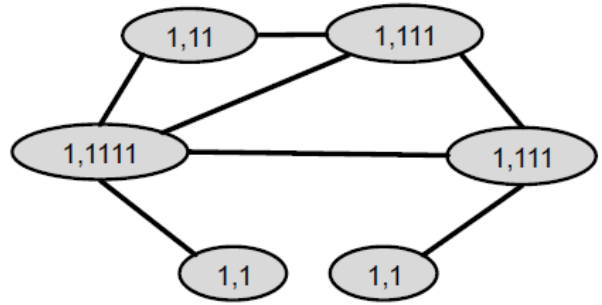


G_2



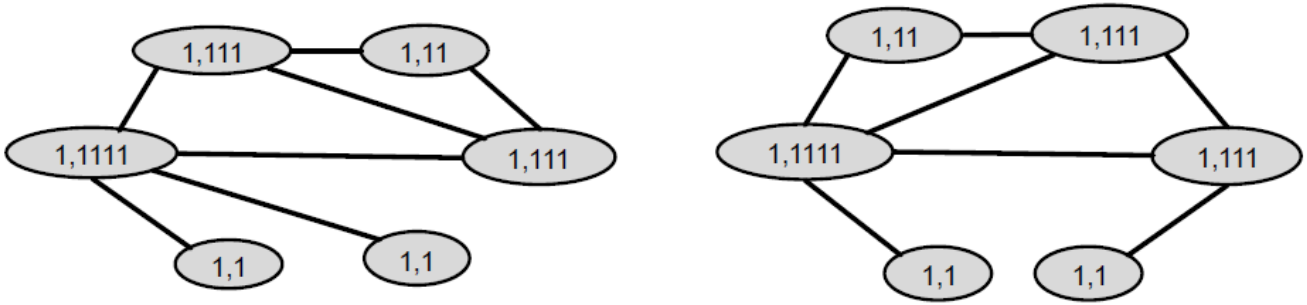
2. Συνένωσε γειτονικά χρώματα:



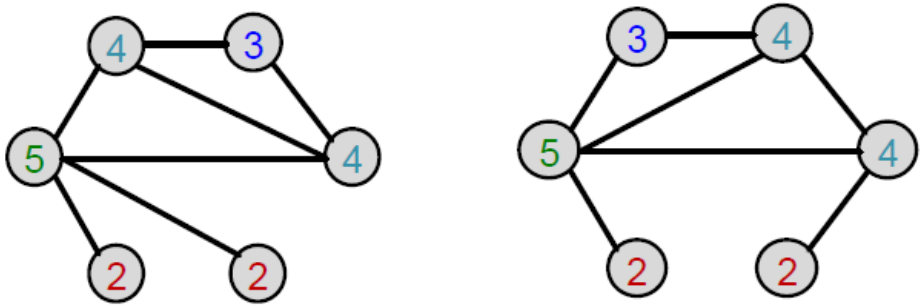


Weisfeiler - Lehman kernel

2. Συνένωσε γειτονικά χρώματα:



3. Κάνε *hashing*:

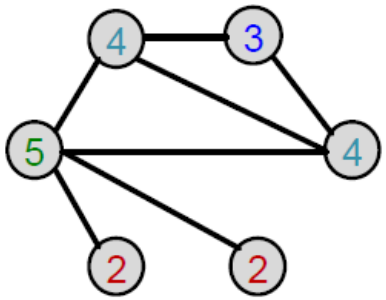


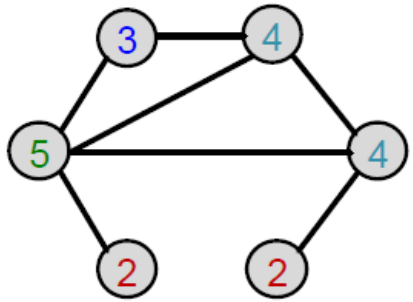
Hash table

1,1	-->	2
1,11	-->	3
1,111	-->	4
1,1111	-->	5

Weisfeiler - Lehman kernel

3. Κάνε *hashing*:

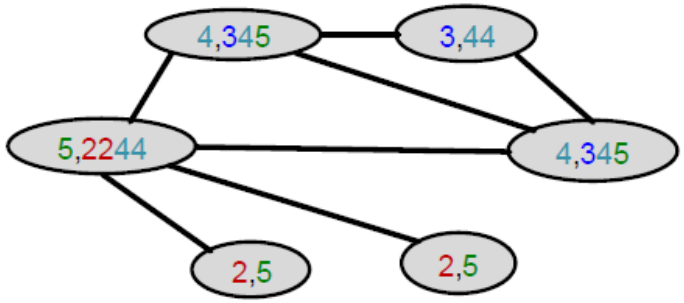


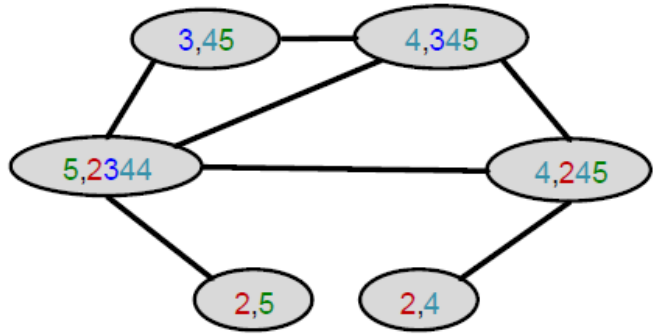


Hash table

1,1	-->	2
1,11	-->	3
1,111	-->	4
1,1111	-->	5

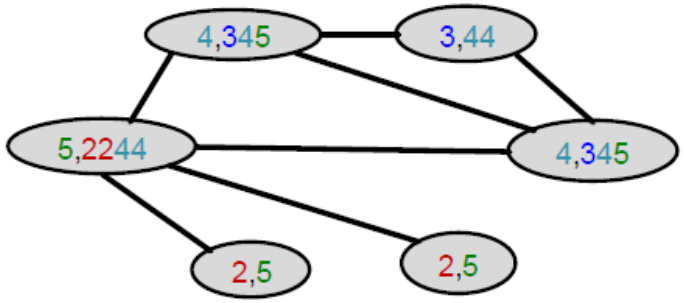
4. Συνένωσε γειτονικά χρώματα:

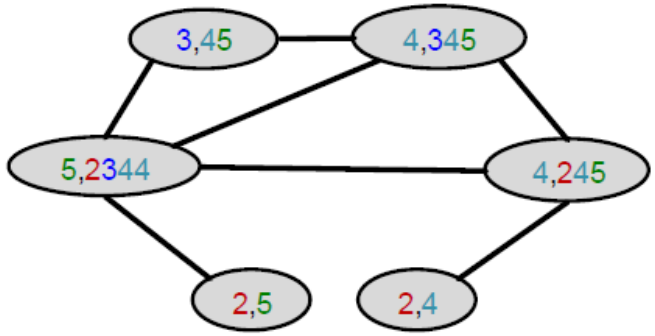




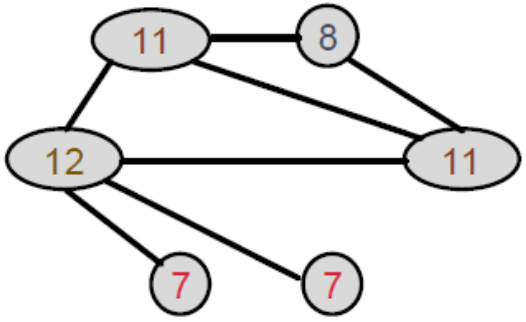
Weisfeiler - Lehman kernel

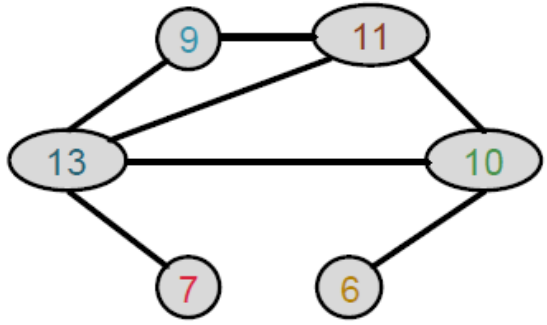
4. Συνένωσε γειτονικά χρώματα:





5. Κάνε *hashing*:



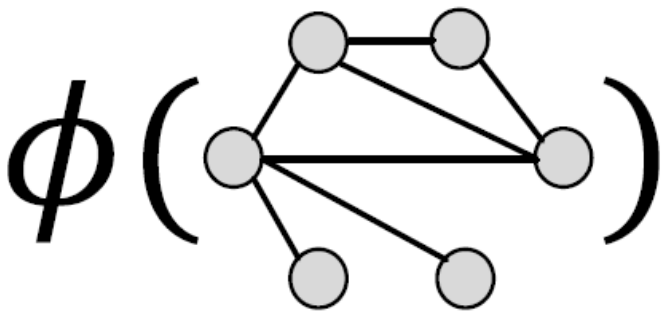


Hash table

2,4	-->	6
2,5	-->	7
3,44	-->	8
3,45	-->	9
4,245	-->	10
4,345	-->	11
5,2244	-->	12
5,2344	-->	13

Weisfeiler - Lehman kernel

Η αναπαράσταση των γραφημάτων έχει ως εξής:

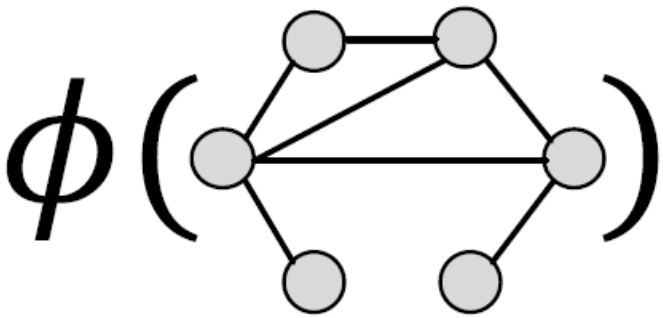


Colors

1,2,3,4,5,6,7,8,9,10,11,12,13

= [6,2,1,2,1,0,2,1,0, 0, 2, 1, 0]

Counts

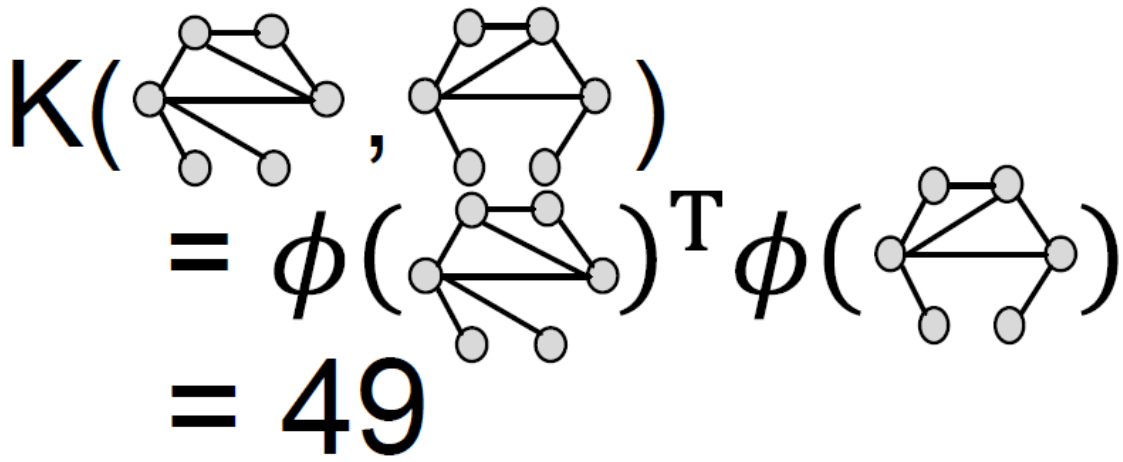


1,2,3,4,5,6,7,8,9,10,11,12,13

= [6,2,1,2,1,1,1,0,1, 1, 1, 0, 1]

Weisfeiler - Lehman kernel

Ο πυρήνας υπολογίζεται από το εσωτερικό γινόμενο ως εξής:



Weisfeiler - Lehman kernel

- Υπολογιστικά αποδοτικός. Ο χρόνος γραμμικός στον αριθμό ακμών διότι συνενώνει γειτονικά χρώματα
- Όταν υπολογίζουμε πυρήνα ο αριθμός χρωμάτων σε 2 γραφήματα είναι στη χειρότερη ίσος με το συνολικό αριθμό κόμβων
- Το μέτρημα των χρωμάτων είναι γραμμικό στον αριθμό κόμβων
- Συνολικό κόστος γραμμικό στον αριθμό των ακμών.

Φασματικές μέθοδοι

- Οι φασματικές μέθοδοι σχετίζονται κυρίως με το πρόβλημα της ομαδοποίησης σε ένα γράφημα
- Δίνουν κίνητρο για μια διανυσματική αναπαράσταση των κόμβων χαμηλής διάστασης
- Ακολουθούν βασικοί ορισμοί

Φασματικές μέθοδοι

- Ο Λαπλασιανός πίνακας L ενός γραφήματος ορίζεται ως $L = DA$, όπου D είναι ο διαγώνιος πίνακας βαθμών και A είναι ο πίνακας γειτνίασης.
- Ιδιοτιμή ενός πίνακα είναι ένα βαθμωτό μέγεθος λ για την οποία υπάρχει ένα μη μηδενικό διάνυσμα (ιδιοδιάνυσμα) τέτοιο ώστε $Av = \lambda v$.
- Η πολλαπλότητα μιας ιδιοτιμής είναι ο αριθμός των γραμμικά ανεξάρτητων ιδιοδιανυσμάτων που σχετίζονται με αυτήν.
- η πολλαπλότητα της μηδενικής ιδιοτιμής του Λαπλασιανού είναι η διάσταση του μηδενικού χώρου του L
- Ένα συνδεδεμένο στοιχείο ενός γραφήματος είναι ένα υπογράφημα στο οποίο οποιεσδήποτε δύο κορυφές συνδέονται μεταξύ τους μέσω μονοπατιών και το οποίο δεν συνδέεται με καμία πρόσθετη κορυφή στο υπεργράφημα.

Φασματικές μέθοδοι: συνδεδεμένα στοιχεία

Το μηδέν είναι μια ιδιοτιμή για κάθε συνδεδεμένο στοιχείο:

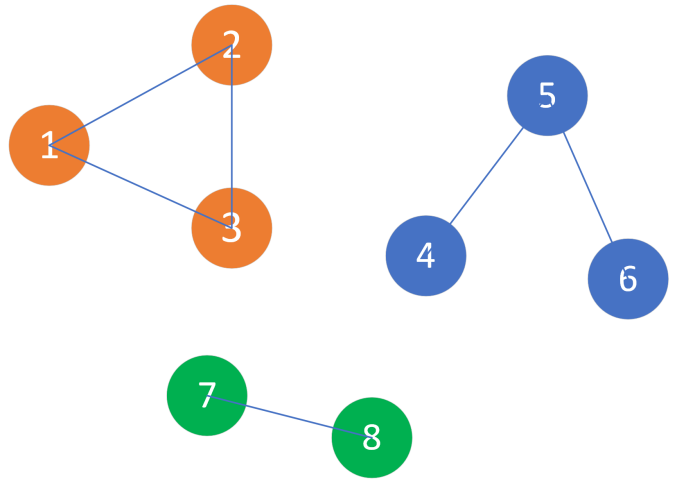
- Θεωρήστε ένα γράφημα G με n κορυφές. Εάν το γράφημα δεν είναι συνδεδεμένο, μπορεί να χωριστεί σε συνδεδεμένα στοιχεία.
- Για κάθε συνδεδεμένο στοιχείο, δημιουργήστε ένα διάνυσμα v μήκους n όπου $v_i = 1$ εάν ο κόμβος i είναι σε αυτό το στοιχείο και $v_i = 0$ διαφορετικά.
- Για κάθε συνδεδεμένο στοιχείο, $Lv = \mathbf{0}$ (καθώς το άθροισμα των βαθμών των κορυφών στο συνδεδεμένο στοιχείο μείον το άθροισμα των ακμών εντός της συνιστώσας είναι μηδέν).
- Άρα, $v^T Lv = 0$ και το v είναι ένα ιδιοδιάνυσμα που σχετίζεται με την ιδιοτιμή 0.

Φασματικές μέθοδοι: συνδεδεμένα στοιχεία

- **Γραμμική ανεξαρτησία ιδιοδιανυσμάτων:**
  Τα ιδιοδιανύσματα που αντιστοιχούν σε διαφορετικά συνδεδεμένα στοιχεία είναι γραμμικά ανεξάρτητα επειδή έχουν μη μηδενικές εγγραφές σε αμοιβαία αποκλειόμενες θέσεις (το καθένα αντιστοιχεί μοναδικά σε ένα συνδεδεμένο στοιχείο).
- **Μετρώντας τα συνδεδεμένα στοιχεία:**
  Ο αριθμός τέτοιων ανεξάρτητων ιδιοδιανυσμάτων που μπορούμε να κατασκευάσουμε είναι ίσος με τον αριθμό των συνδεδεμένων στοιχείων στο γράφημα. Αυτό συμβαίνει επειδή κάθε συνδεδεμένο στοιχείο συνεισφέρει ακριβώς ένα ανεξάρτητο ιδιοδιάνυσμα στον μηδενικό χώρο του L.
- **Η πολλαπλότητα ισούται με τον αριθμό των στοιχείων:**
  Εφόσον η διάσταση του μηδενικού χώρου του L είναι ίση με τον αριθμό αυτών των ανεξάρτητων ιδιοδιανυσμάτων, προκύπτει ότι η πολλαπλότητα της μηδενικής ιδιοτιμής του L είναι ίση με τον αριθμό των συνδεδεμένων στοιχείων.

Φασματικές μέθοδοι: συνδεδεμένα στοιχεία

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$



$$L = D - A = \begin{pmatrix} -2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Φασματικές μέθοδοι: συνδεδεμένα στοιχεία

Αν $x(u)$ το στοιχείο του διανυσματος x στην αντίστοιχη θέση του κόμβου u ισχύει ότι:

$$x^T Lx = \frac{1}{2} \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} A[u, v](x(u) - x(v))^2$$

ή αλλιώς:

$$x^T Lx = \sum_{(u,v) \in \mathcal{E}} (x(u) - x(v))^2$$

Στην τετραγωνική μορφή θα εμφανίζονται για κάθε ακμή (u, v) οι εξής όροι του αθροίσματος:

- $L_{u,v}x(u)x(v)$ και $L_{v,u}x(u)x(v)$ $(L_{u,v} = L_{v,u} = -1)$ για κάθε ακμή, δηλαδή το αθροισμα τους θα είναι $-2ex(u)x(v)$
- $x(u)^2$, $x(v)^2$

Βάσει αυτών προκύπτει ο όρος του αθροίσματος $(x(u) - x(v))^2$

Φασματικές μέθοδοι: συνδεδεμένα στοιχεία

Σύμφωνα με τα παραπάνω για το παράδειγμα ισχύει:
$v_1 = (1, 1, 1, 0, 0, 0, 0, 0)^T$
$v_2 = (0, 0, 0, 1, 1, 1, 0, 0)^T$
$v_3 = (0, 0, 0, 0, 0, 0, 1, 1)^T$

Βρίσκουμε εύκολα ότι $L \cdot v_i = \mathbf{0} = 0 \cdot v_i$ άρα τα v_i είναι ιδιοδιανύσματα που αντιστοιχούν σε μηδενική ιδιοτιμή.

Επίσης $v_i \cdot v_j = 0$, $i \neq j$, άρα γραμμικά ανεξάρτητα.

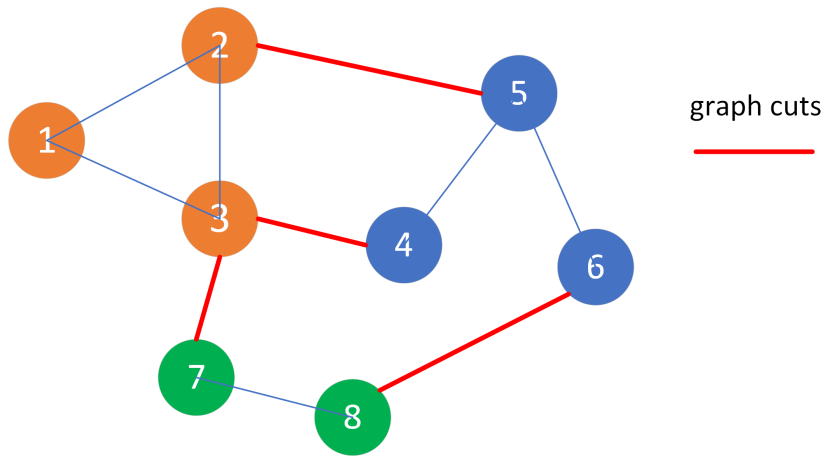
Η πολλαπλότητα της μηδενικής ιδιοτιμής είναι 3, όσο και ο αριθμός των συνδεδεμένων στοιχείων.

Οι ιδιοτιμές του L δίνονται από την ένωση των συνόλων των ιδιοτιμών των συνδεδεμένων στοιχείων.

Graph cuts

Έστω ότι έχουμε k ομάδες κόμβων. Ορίζουμε graph cut:

$$cut(\mathcal{A}_1, ..., \mathcal{A}_K) = \frac{\sum_{k=1}^{K} \left|(u,v) \in \mathcal{E} : u \in \mathcal{A}_k, u \in \mathcal{A}_k^c\right|}{2}$$



k=3, $cut(\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3) = 4$

Graph cuts

Προσπαθούμε να βρούμε την ομαδοποίηση έτσι ώστε να ελαχιστοποιήσουμε το $cut(\mathcal{A}_1, ..., \mathcal{A}_K)$.

Πιο καλές λύσεις βρίσκουμε όταν προσπαθούμε να ελαχιστοποιήσουμε το

$$RatioCut(\mathcal{A}_1, ..., \mathcal{A}_K) = \frac{\sum_{k=1}^{K} \left|(u,v) \in \mathcal{E} : u \in \mathcal{A}_k, u \in \mathcal{A}_k^c\right|}{2|\mathcal{A}_k|}$$

Για K=2 το ιδιοδιάνυσμα που αντιστοιχεί στη δεύτερη μικρότερη ιδιοτιμή της Λαπλασιανής είναι μια συνεχής προσέγγιση στο διακριτό διάνυσμα που δίνει μια βέλτιστη κατανομή στις 2 ομάδες (βλ. Hamilton 2020).

Η ιδέα μπορεί να επεκταθεί σε k ομάδες λαμβάνοντας κάθε φορά υπόψη το ιδιοδιάνυσμα της k - μικρότερης ιδιοτιμής.

Βιβλιογραφία

1. W.L. Hamilton, Graph Representation Learning, McGill University, 2020
2. J. Leskovec, Machine Learning with Graphs, Stanford University, Fall 2023