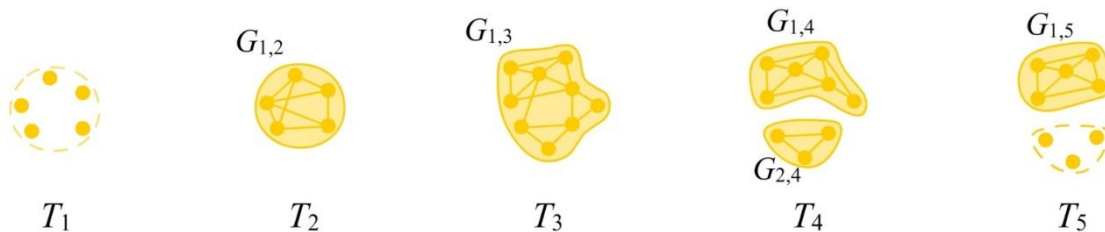
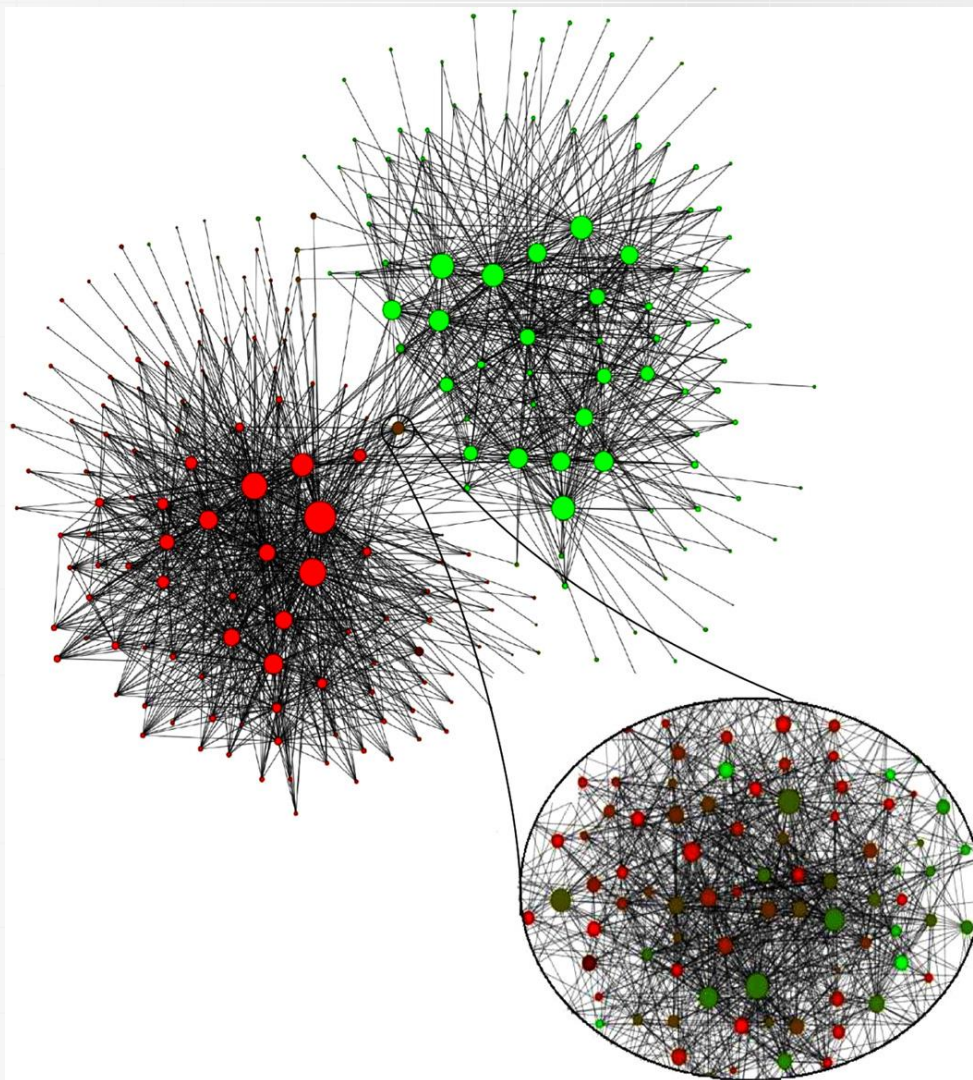


Sieci złożone

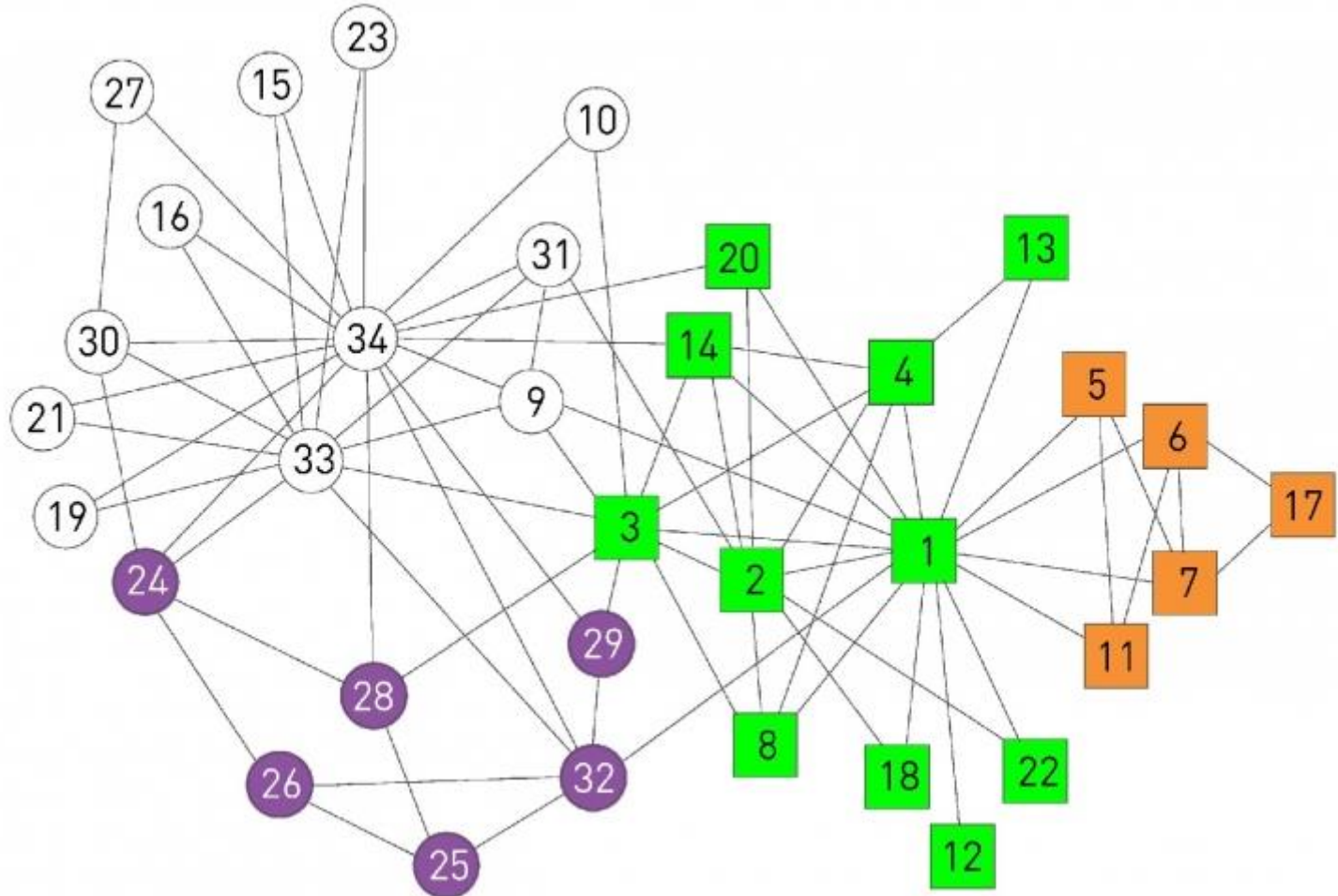


Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

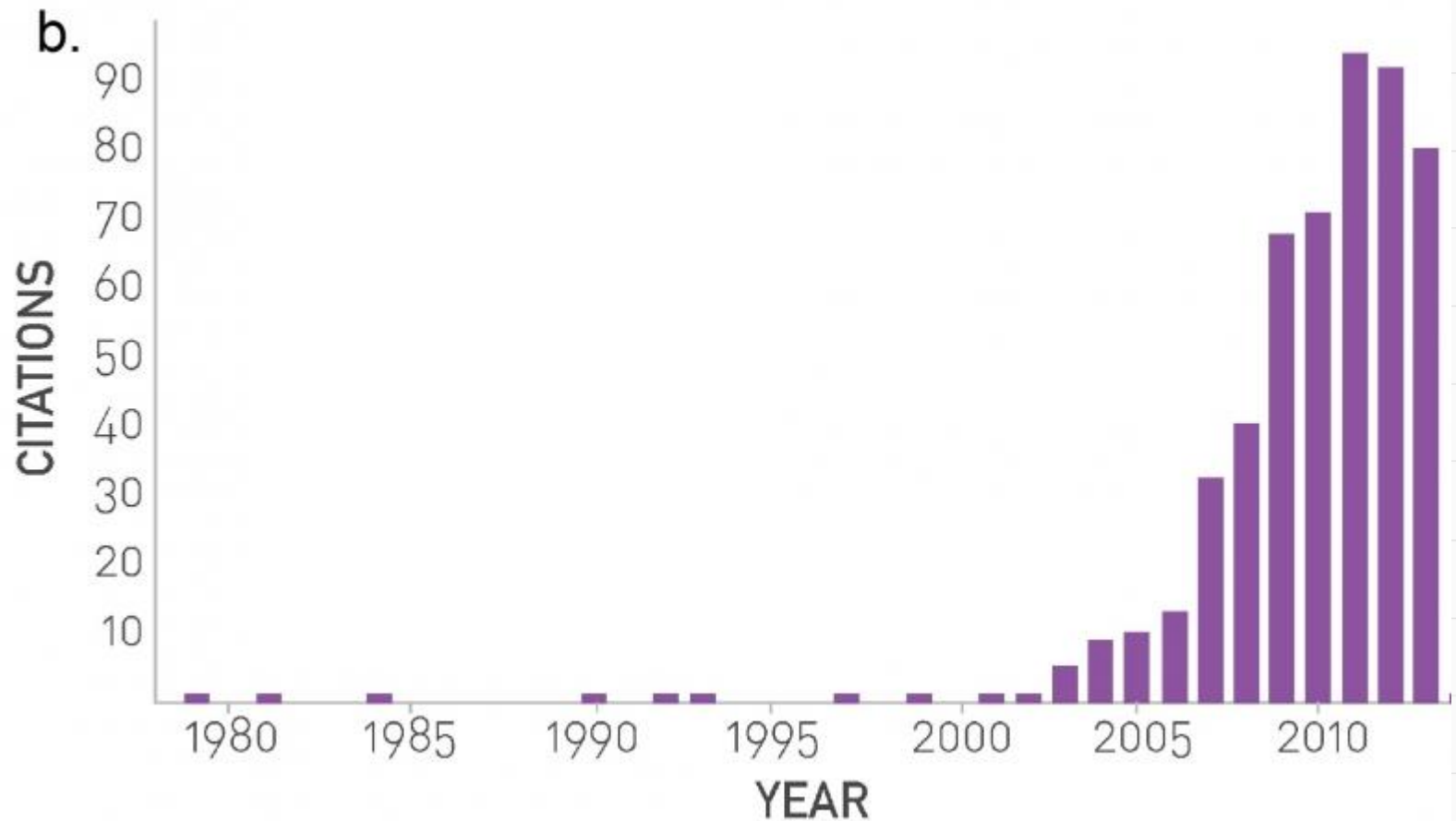
Podzieleni



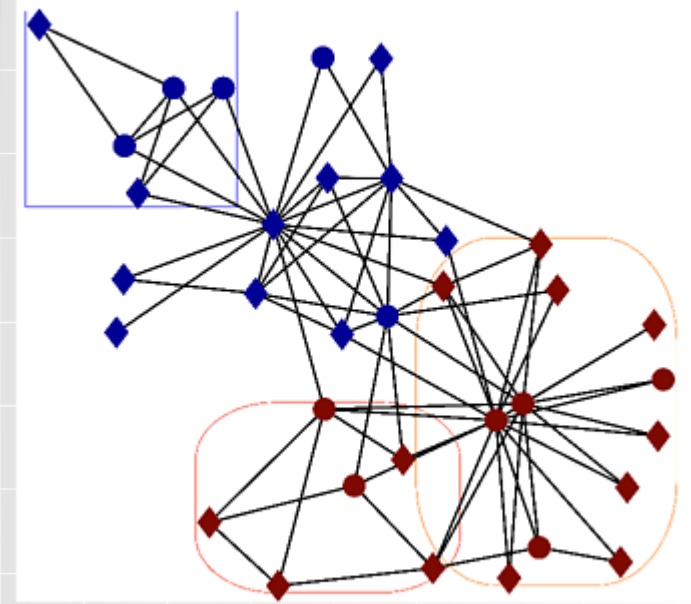
Zachary's Karate Club



Zachary's Karate Club



<http://networkkarate.tumblr.com/>

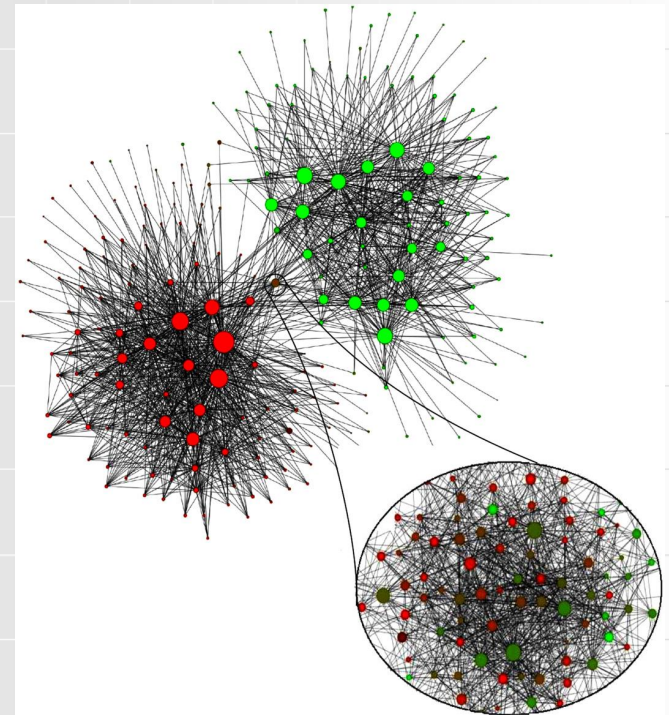
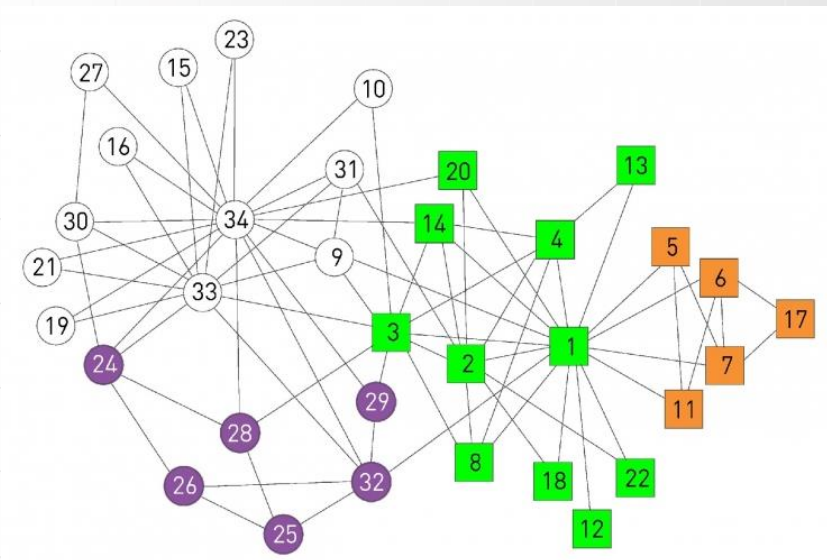


If you can't get it right on this network, then go home.



Uzupełniająca wiedza

- Klub Karate – Konflikt i rozpad klubu



- Sieć połączeń telefonicznych w Belgii – struktura społeczeństwa a w szczególności język.



Politechnika
Wrocławska

Podstawy wykrywania grup



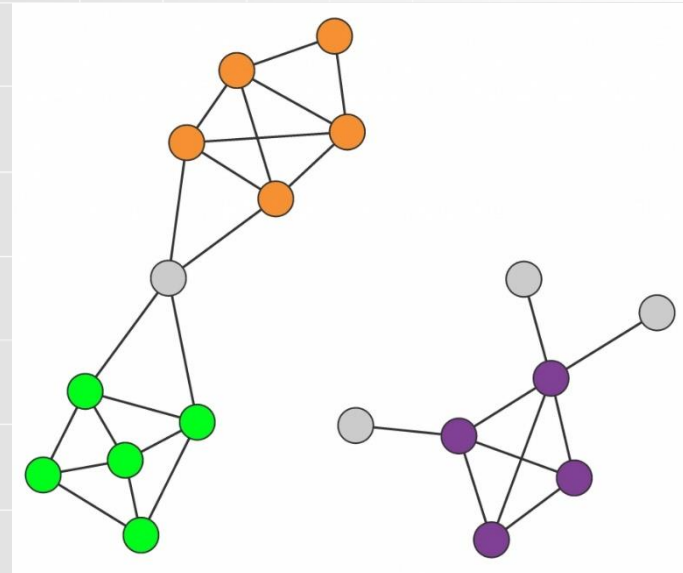
HR EXCELLENCE IN RESEARCH

Podstawy wykrywania grup

Grupy (groups, communities) to lokalnie gęsto połączone podgrafy w sieci. To oczekiwanie opiera się na dwóch odrębnych hipotezach:

- **Hipoteza łączności** – grupa jest połączonym grafem
- **Hipoteza gęstości** – grupa jest lokalnie gęstym podgrafem sieci

$$\text{gęstość/density} = D = \delta = \frac{2L}{N(N-1)}$$



Podstawy wykrywania grup

- **Stopień wewnętrzny** (Internal degree), k^{int} : liczba krawędzi która łączy wierzchołki danej grupy.
 - Jeżeli $k_i^{int} = 0$, wierzchołek i nie powinien należeć do grupy
- **Stopień zewnętrzny** (External degree) k^{ext} : liczba krawędzi która łączy wierzchołki danej grupy z pozostałymi wierzchołkami w sieci.
 - Jeżeli $k_i^{ext} = 0$, dana grupa jest najlepszą dla wierzchołka i
- Gęstość wewnętrzna grupy
intra-cluster density
- Gęstość zewnętrzna grupy
inter-cluster density

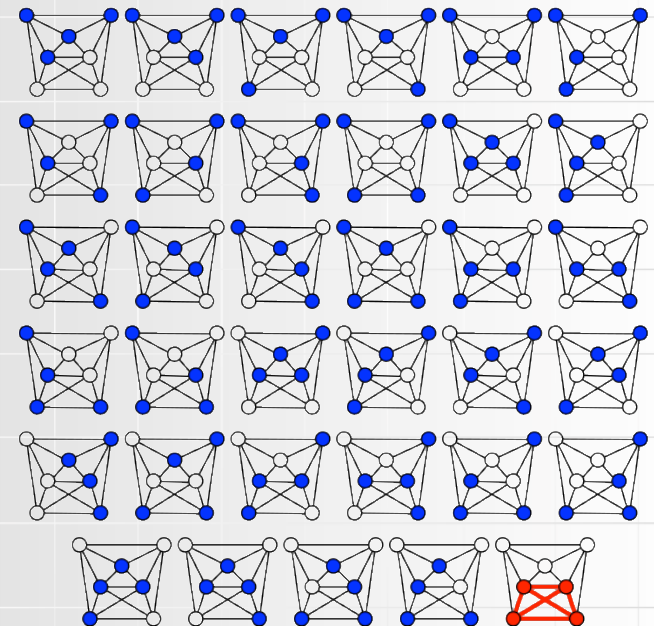
$$\delta^{int}(C) = \frac{k^{int}}{N_c(N_c - 1)}$$

$$\delta^{ext}(C) = \frac{k^{ext}}{N_c(N - N_c)}$$

$$\delta^{int}(C) \gg \langle k \rangle \gg \delta^{ext}(C)$$

Definicja grupy

- Klia
 - Podgraf będący grafem pełnym
 $\delta^{\text{int}}(C) = 1$
 - Trójkąty (3-kliki, 3-clique)
występują często, większe kliki
są rzadkie
 - W rzeczywistości grupy nie
muszą być grafami pełnymi
 - Wyszukiwanie klik jest
problemem NP.-trudnym



Definicja grupy

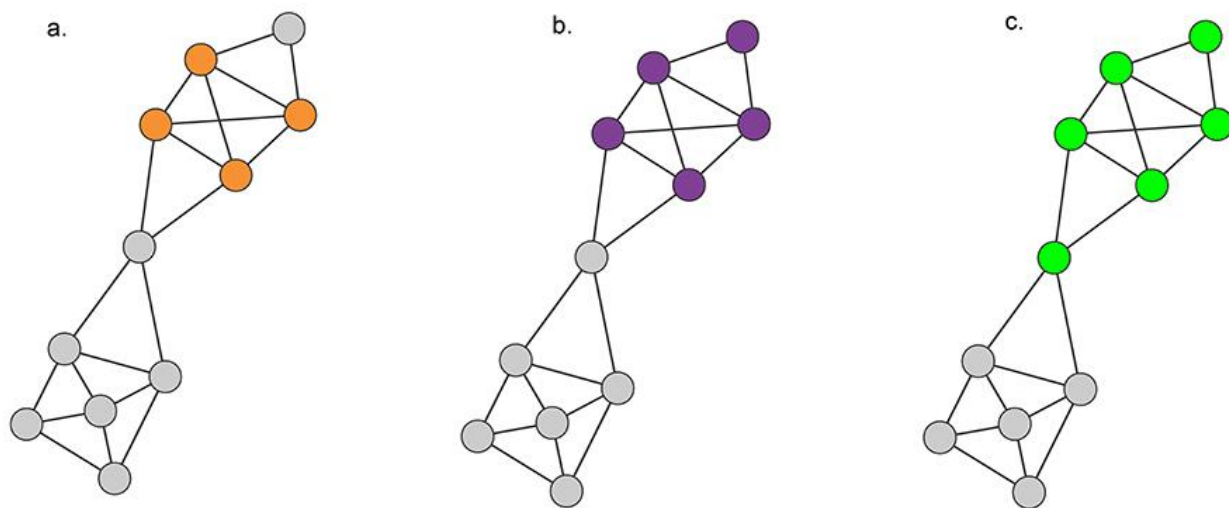
- **Silna grupa** - każdy węzeł ma więcej krawędzi łączących go z wierzchołkami wewnątrz grupy niż z wierzchołkami poza grupą.

$$k_i^{int}(C) > k_i^{ext}(C)$$

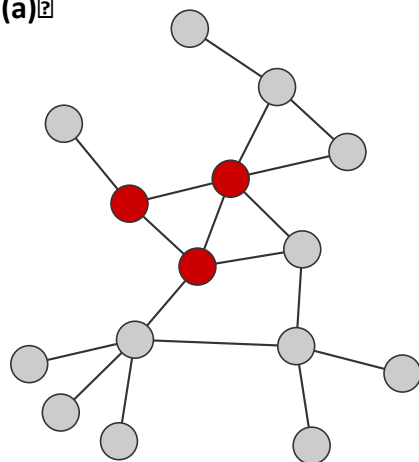
- **Słaba grupa** - całkowity stopień wewnętrzny grupy przekracza jego całkowity stopień zewnętrzny grupy

$$\sum_{i \in C} k_i^{int}(C) = k^{int}(C) > k^{ext}(C) = \sum_{i \in C} k_i^{ext}(C)$$

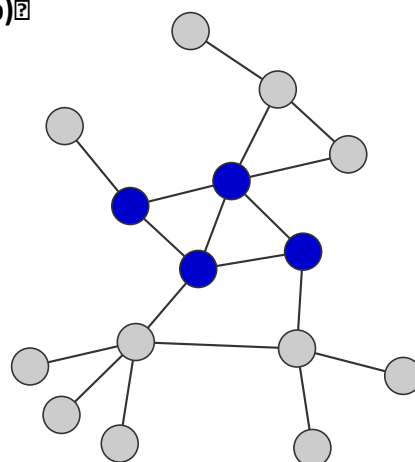
Definicja grupy



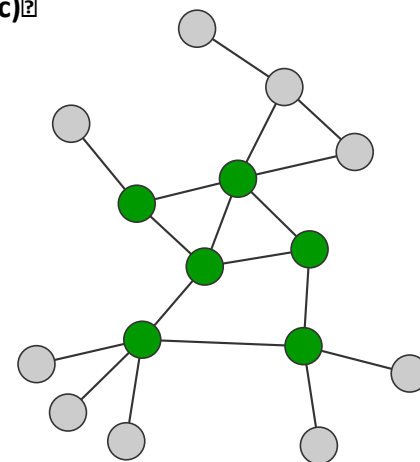
(a)



(b)



(c)





Politechnika
Wrocławska

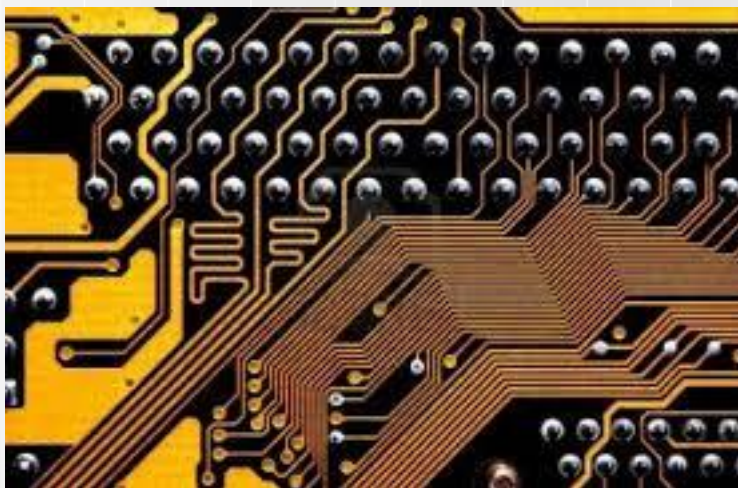
Partycjonowanie grafu a wykrywanie grup



HR EXCELLENCE IN RESEARCH

Partycjonowanie grafu

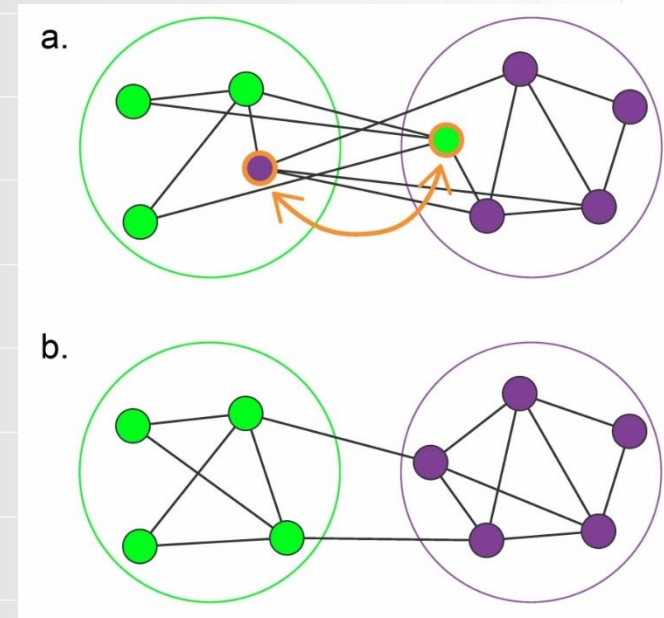
- Procesor - miliardy tranzystorów
 - Partycjonowanie – taki podział/ułożenie tranzystorów na partycje na płycie by zminimalizować liczbę połączeń między nimi.



Kernighan, B. W., & Lin, S. (1970). **An efficient heuristic procedure for partitioning graphs.** *The Bell system technical journal*, 49(2), 291-307.

Algorytm Kernighan-Lin

- Podziel sieć na dwie grupy o predefiniowanym rozmiarze.
- Sprawdź każdą parę węzłów, po jednym z każdej grupy. Zidentyfikuj parę, która jeżeli zostanie zamieniona między partycjami powoduje największe zmniejszenie liczby krawędzi do przecięcia.
- Jeżeli żadna para nie powoduje zmniejszenia liczby krawędzi do przecięcia, zamień tę, która powoduje najmniejszy wzrost.
- Zamień wierzchołki miejscami
- Proces jest powtarzany, aż każdy węzeł zostanie przesunięty raz.

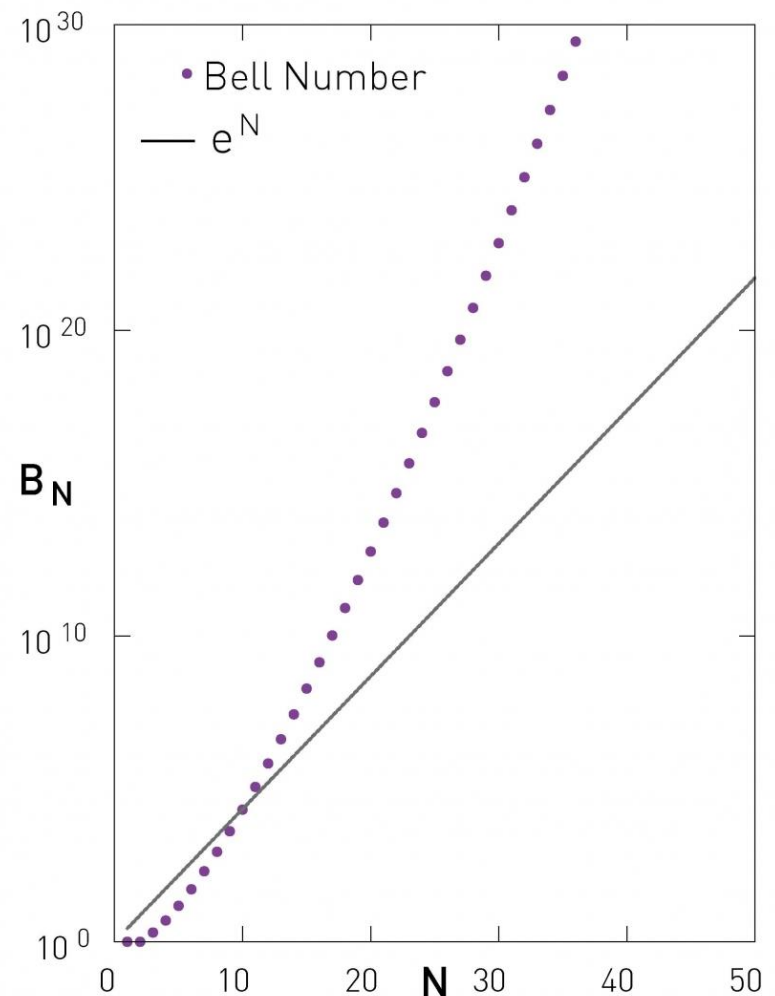


Partycjonowanie grafu a wykrywanie grup

- Partycjonowanie grafów dzieli sieć na określoną liczbę mniejszych podgrafów
- Wykrywanie grup ma na celu odkrycie wewnętrznej struktury społeczności sieci.

Partycjonowanie grafu a wykrywanie grup

- Liczba możliwych sposobów podziału sieci na grupy rośnie wykładniczo lub szybciej wraz z rozmiarem sieci N .
- Dlatego niemożliwe jest sprawdzenie wszystkich partycji dużej sieci
- $N=10 \rightarrow 252$ podziały (1 ms) $\frac{N!}{N_1! N_2!}$
- $N=100 \rightarrow 10^{29}$ podziały (10^{16} lat)



Grupowanie hierarchiczne

- Aby wykryć grupy w dużych sieci rzeczywistych, potrzebujemy algorytmów, których czas działania rośnie liniowo lub wielomianowo wraz ze wzrostem N
- Grupowanie hierarchiczne bazuje na macierzy podobieństwa (**similarity matrix**) gdzie x_{ij} mówi nam jak i jest podobne do j
- Dwa główne podejścia
 - Algorytmy aglomeracyjne (**agglomerative a.**) scalają węzły o dużym podobieństwie w tę samą grupę
 - Algorytmy deglomeracyjne (**divisive a.**) izolują grupy, usuwając linki pomiędzy węzłami o niskim podobieństwie, które zwykle łączą grupy

Aglomeracyjne - algorytm Ravasz

Złożoność obliczeniowa $O(N^2)$

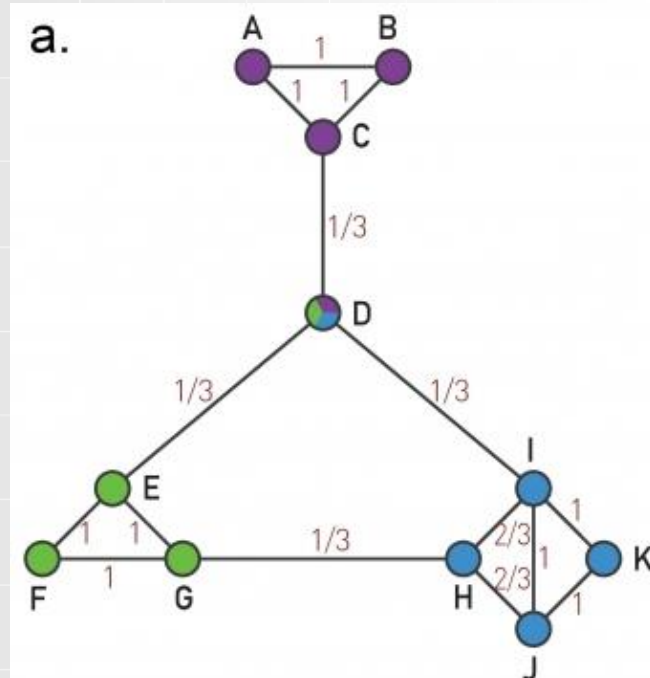
1. Zdefiniuj macierz podobieństwa.

- wysokie dla par węzłów, które prawdopodobnie należą do tej samej grupy;
- niskie dla par, które prawdopodobnie należą do różnych grup.
- Węzły, które łączą się bezpośrednio ze sobą i/lub współdzielą wielu sąsiadów, z większym prawdopodobieństwem należą do tego samego gęstego lokalnego sąsiedztwa, dlatego powinny mieć duże podobieństwo.

Aglomeracyjne - algorytm Ravasz

1. Zdefiniuj macierz podobieństwa.

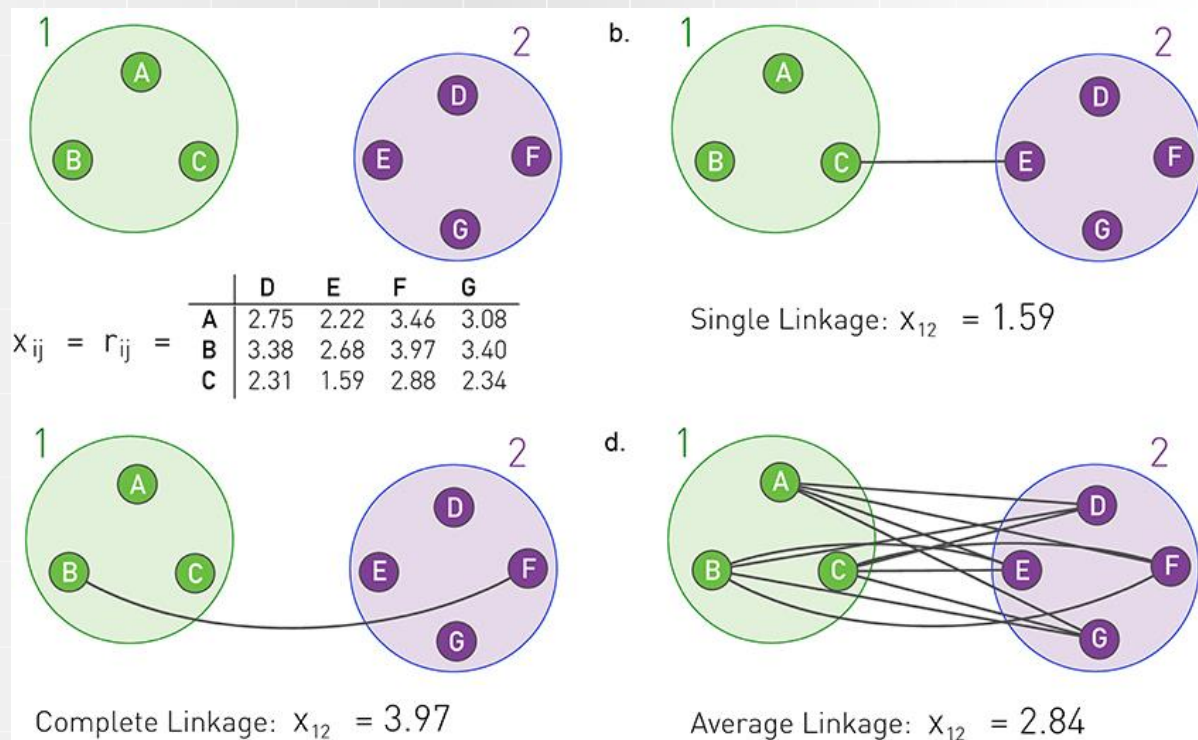
- $$x_{ij} = \frac{J(i,j)}{\min(k_i, k_j) + 1 - \theta(A_{ij})} = \frac{J(i,j)}{\min(k_i, k_j)}$$
 - $J(i, j)$ – liczba wspólnych sąsiadów + 1 jeżeli i i j są połączone
 - $\theta(y)$ – funkcja skokowa Heaviside'a która przyjmuje wartość 1 dla $y > 0$ i 0 dla $y \leq 0$
 - $1 - \theta(A_{ij}) = 0$ dla wierzchołków połączonych
- $x_{ij} = 1$ jeżeli i i j są połączone i mają tych samych sąsiadów
- $x_{ij} = 0$ jeżeli i i j nie są połączone i nie mają wspólnych sąsiadów



Aglomeracyjne - algorytm Ravasz

2. Wybierz jak określisz podobieństwo dwóch grup

- Minimalna odległość (Single Linkage)
- Maksymalna odległość (Complete Linkage)
- Średnia odległość (Average Linkage)



Aglomeracyjne - algorytm Ravasz

3. Wykonaj grupowanie hierarchiczne

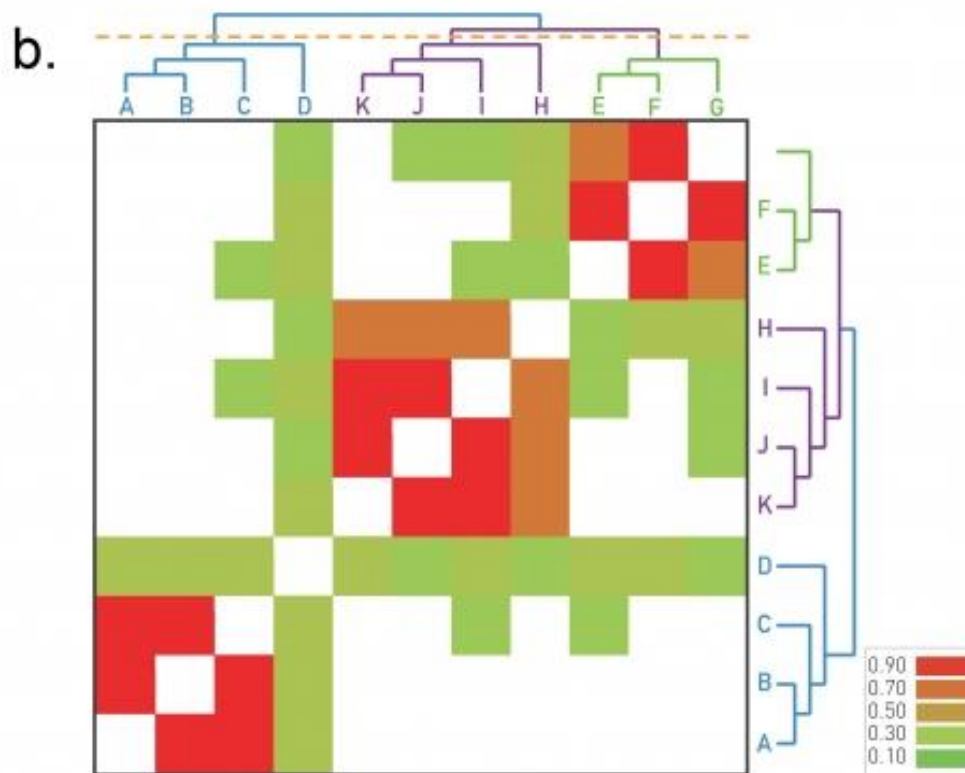
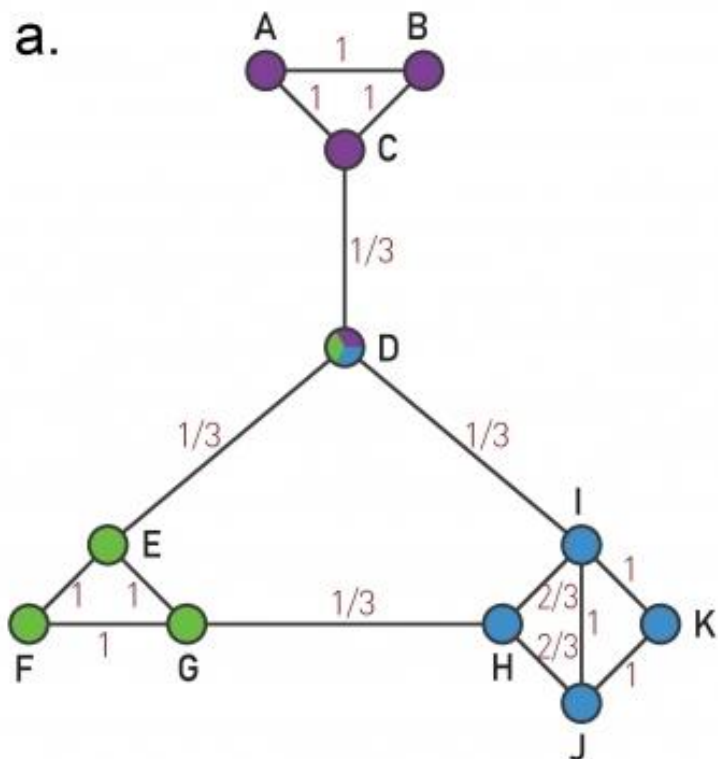
- 1) Przypisz każdy węzeł do „własnej” grupy i oceń podobieństwo dla wszystkich par węzłów. Początkowe podobieństwa między „grupami” to po prostu podobieństwo pomiędzy węzłami.
- 2) Znajdź parę grup o największym podobieństwie i połącz je w jedną grupę.
- 3) Oblicz podobieństwo między nową grupą a wszystkimi innymi grupami.
- 4) Powtarzaj od kroku 2), aż wszystkie węzły zostaną połączone w jedną grupę.

4. Stwórz dendogram

- opisuje dokładną kolejność, w jakiej węzły są przypisywane do grup.

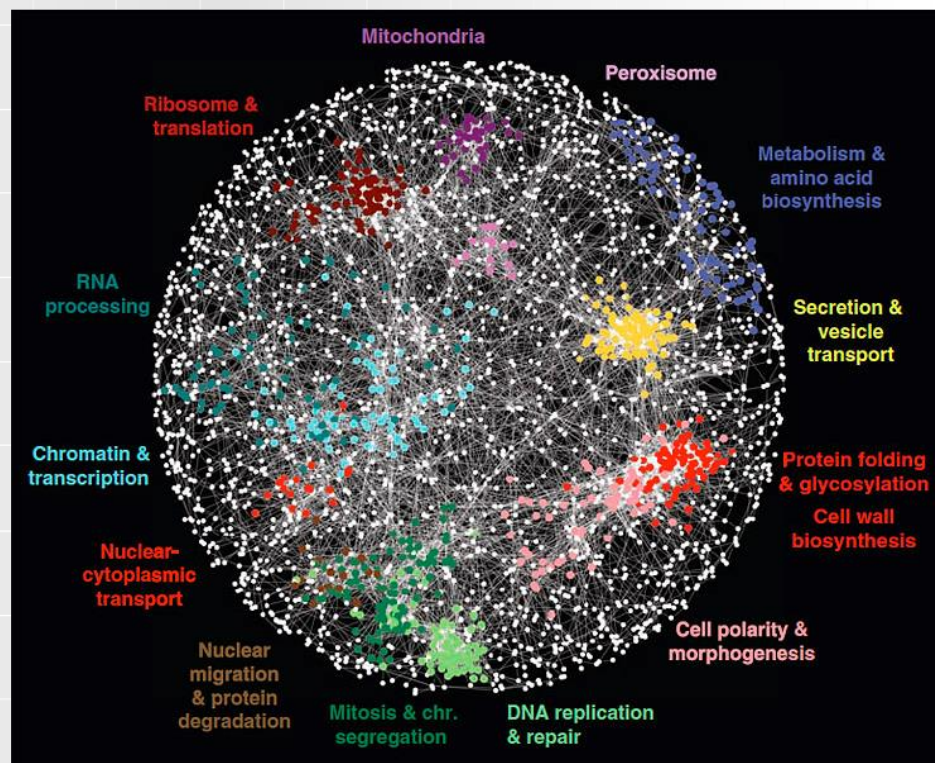
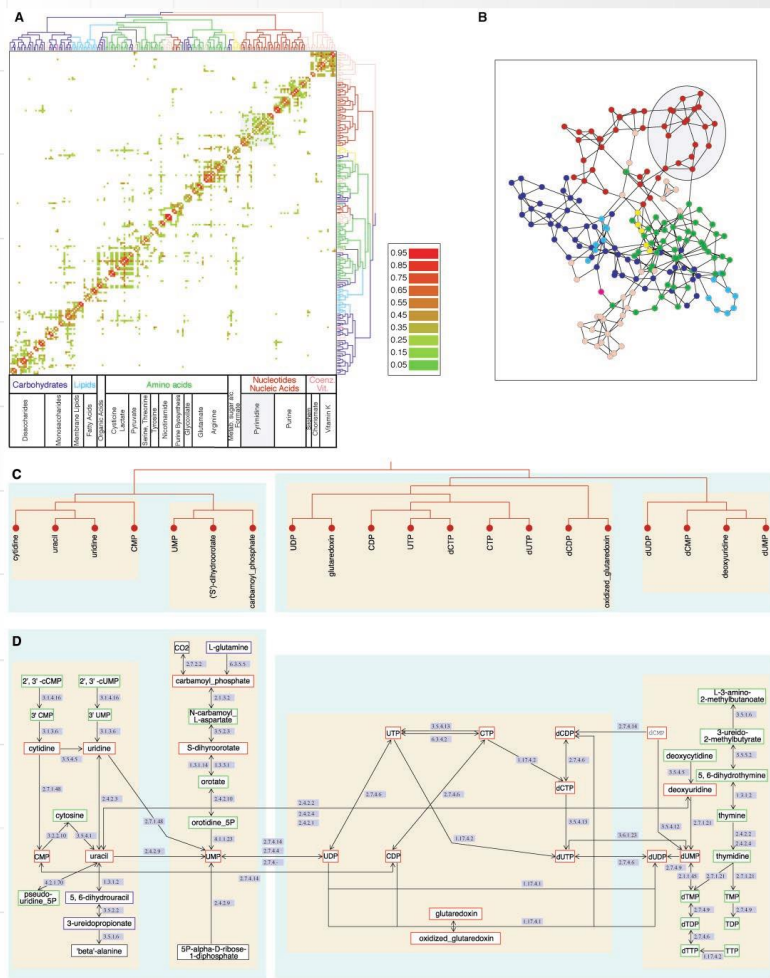
5. Wybierz punkt przecięcia

Aglomeracyjne - algorytm Ravasz



Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551-1555.

Grupy w sieci metabolicznej



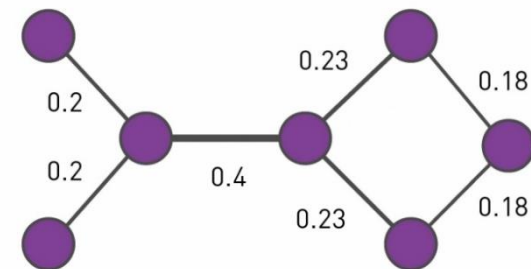
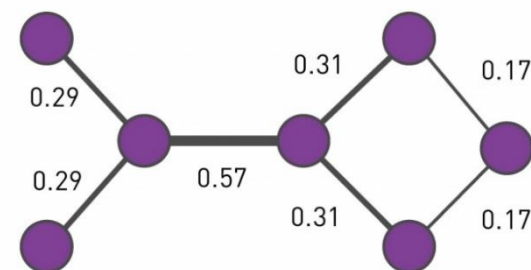
Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

Deglomeracyjne: alg. Girvan – Newman

- Usuwać krawędzie łączące węzły należące do różnych społeczności, dzięki czemu sieć będzie się dzielić na izolowane grupy.

1. Wybierz miarę centralności określającą bliskość węzłów

- **Przewodnictwo dla krawędzi (Link betweenness)**
- Współczynnik grupowania dla krawędzi (Edge clustering coefficient)
- Błądzenie losowe dla przewodnictwa (Random-walk link betweenness)
- itd.



Girvan, M., & Newman, M. E. (2002). **Community structure in social and biological networks.** *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

Deglomeracyjne: alg. Girvan – Newman

2. Wykonaj grupowanie hierarchiczne

- 1) Oblicz podobieństwo x_{ij} dla każdej krawędzi.
- 2) Usuń krawędź z największą wartością x_{ij} . Jeżeli dwie krawędzie mają tę samą wartość x_{ij} usuń losowo jedną z nich.
- 3) Oblicz ponownie podobieństwo x_{ij} dla każdej krawędzi w nowej sieci
- 4) Powtórz od kroku 2) aż usuniesz wszystkie krawędzie.

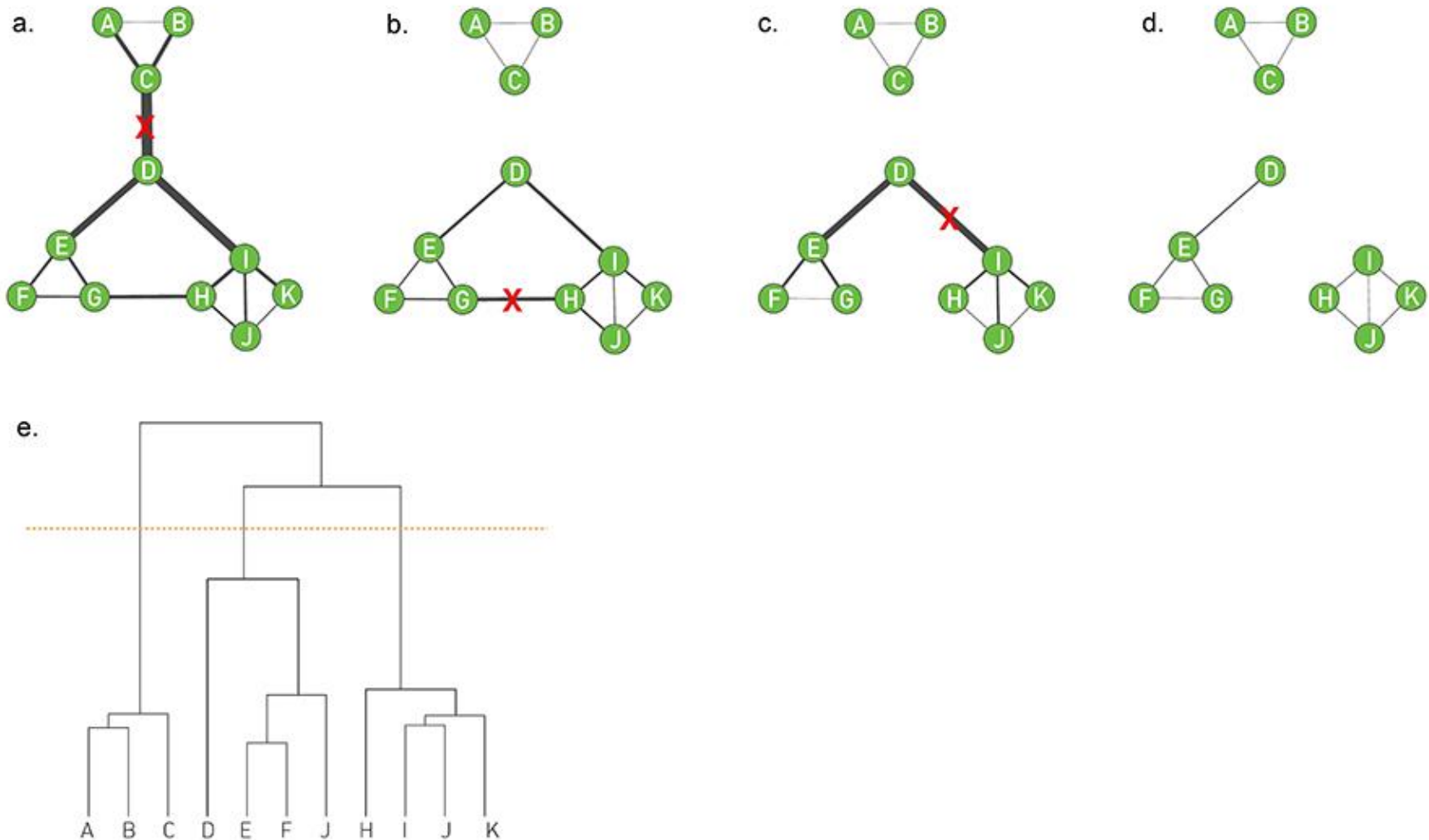
3. Stwórz dendogram

4. Wybierz punkt przecięcia

Złożoność obliczeniowa $O(LN)$, $O(N^2)$ dla sieci rzadkich

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

Deglomeracyjne: alg. Girvan – Newman



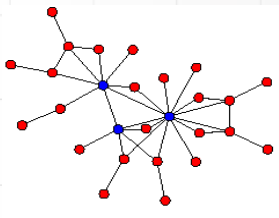
Grupowanie hierarchiczne - problemy

- Założenie że istnieje hierarchia:

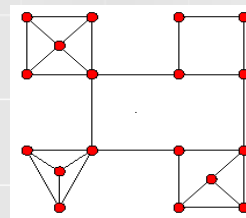
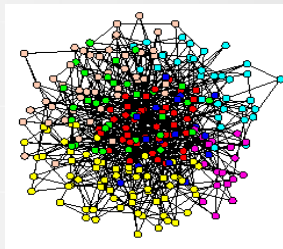
Grupy są zorganizowane w sposób hierarchiczny. Ta hierarchia może być uchwycona przez dendrogram, pokazujący jak mniejsze grupy są zagnieżdżone w większych.

- Związek z sieciami bezskalowymi:

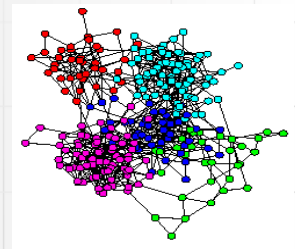
Wykrywanie grup zakłada, że sieć może zostać podzielona na zbiór podgrafów, które są lokalnie gęste i są tylko w niewielkim stopniu połączone z innymi grupami. Jednak w sieciach bezskalowych są węzły, które ze względu na wysoki stopień wężła mogą łączyć się z węzłami należącymi do różnych grup.



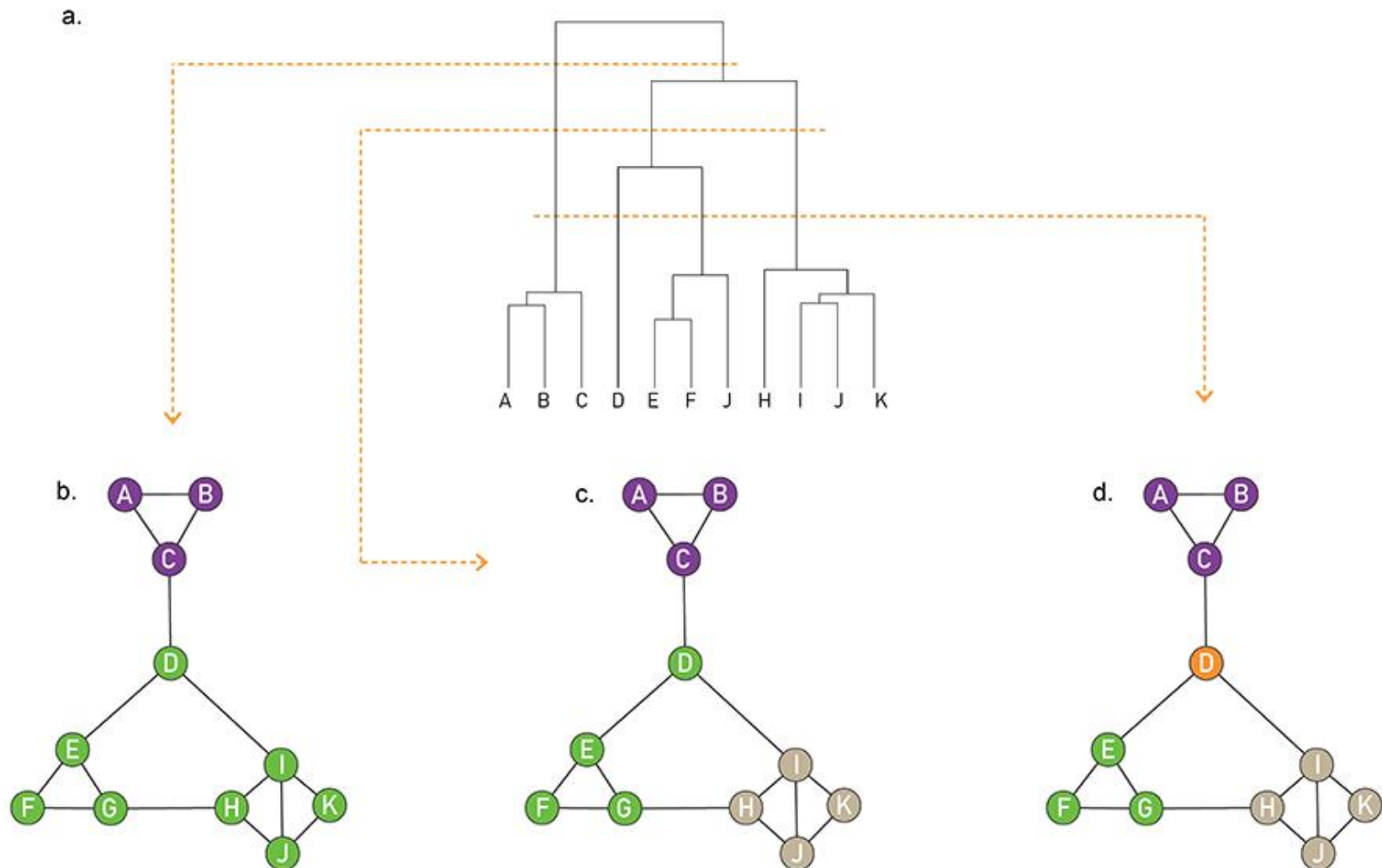
Scale-free



Modular



Niejednoznaczność w punkcie wyborze punktu przecięcia





Politechnika
Wrocławska

Modularność (Modularity)



HR EXCELLENCE IN RESEARCH

Modularność

U podstawy modularności leży założenie że w sieci losowej rozkład połączeń między węzłami powinien być jednorodny i niezależny od rozkładu stopni węzła. W konsekwencji w sieciach losowych* nie oczekujemy znalezienia lokalnie gęstych podgrafów które moglibyśmy interpretować jako grupy.

*np. tu nam się przydają sieci losowe :)

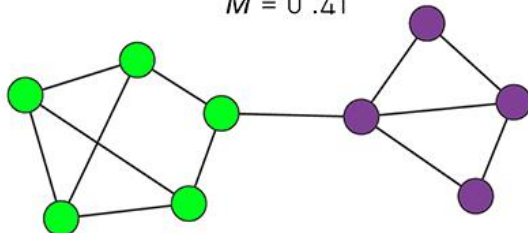
Modularność grupy

- $M_c = \frac{1}{L} \sum_{(i,j) \in C_c} (A_{ij} - p_{ij})$
- $p_{ij} = \frac{k_i k_j}{2L}$ używamy modelu zerowego (null model) z zachowaniem stopnia wężła (sieć losowa)
- $M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2$
- C_c - grupa (podgraf)
- L_c - liczba krawędzi w grupie C_c
- k_c - suma stopni wężła, wężłków w grupie C_c

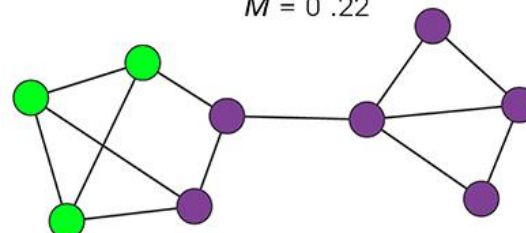
Modularność sieci

$$M = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

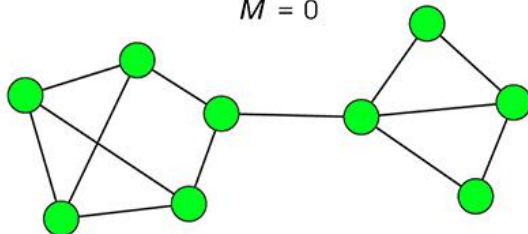
a. OPTIMAL PARTITION
 $M = 0.41$



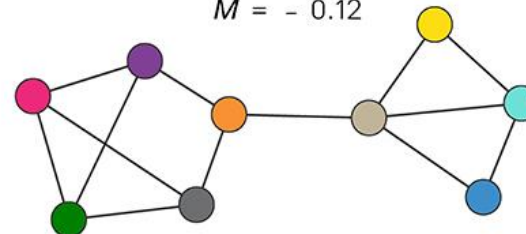
b. SUBOPTIMAL PARTITION
 $M = 0.22$



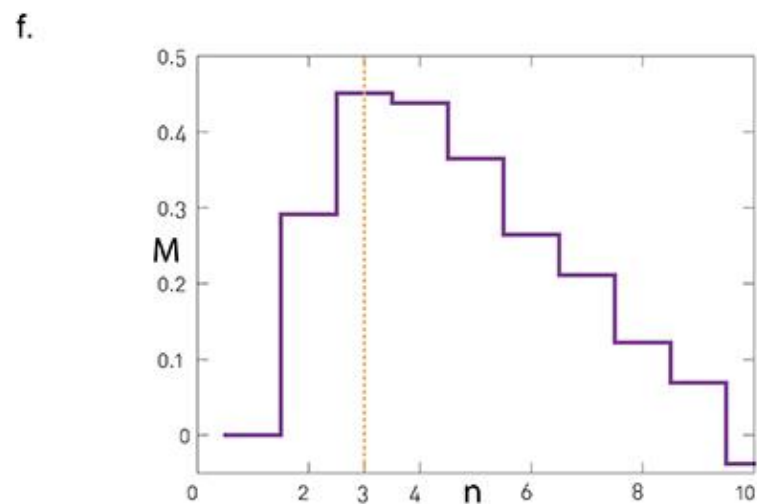
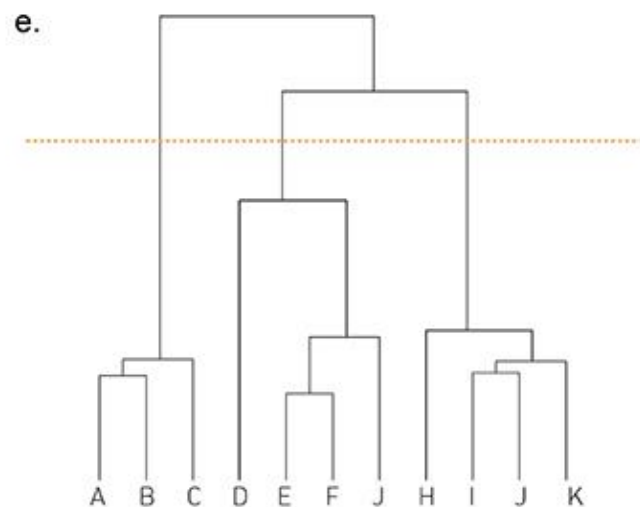
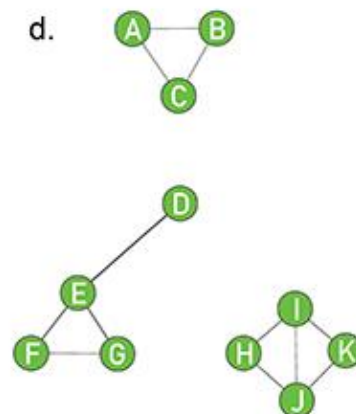
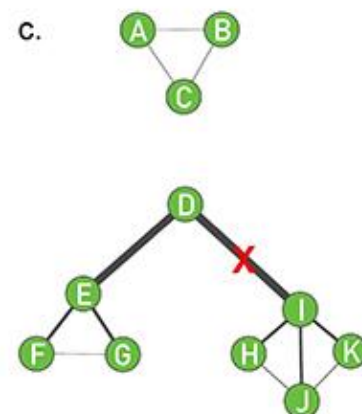
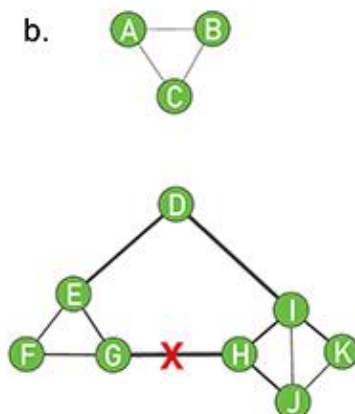
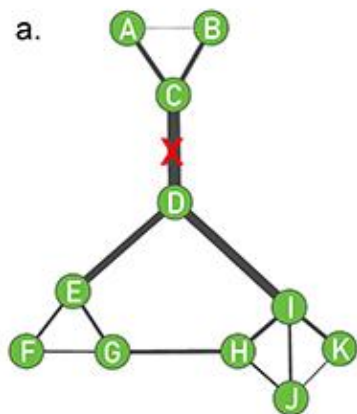
c. SINGLE COMMUNITY
 $M = 0$



d. NEGATIVE MODULARITY
 $M = -0.12$



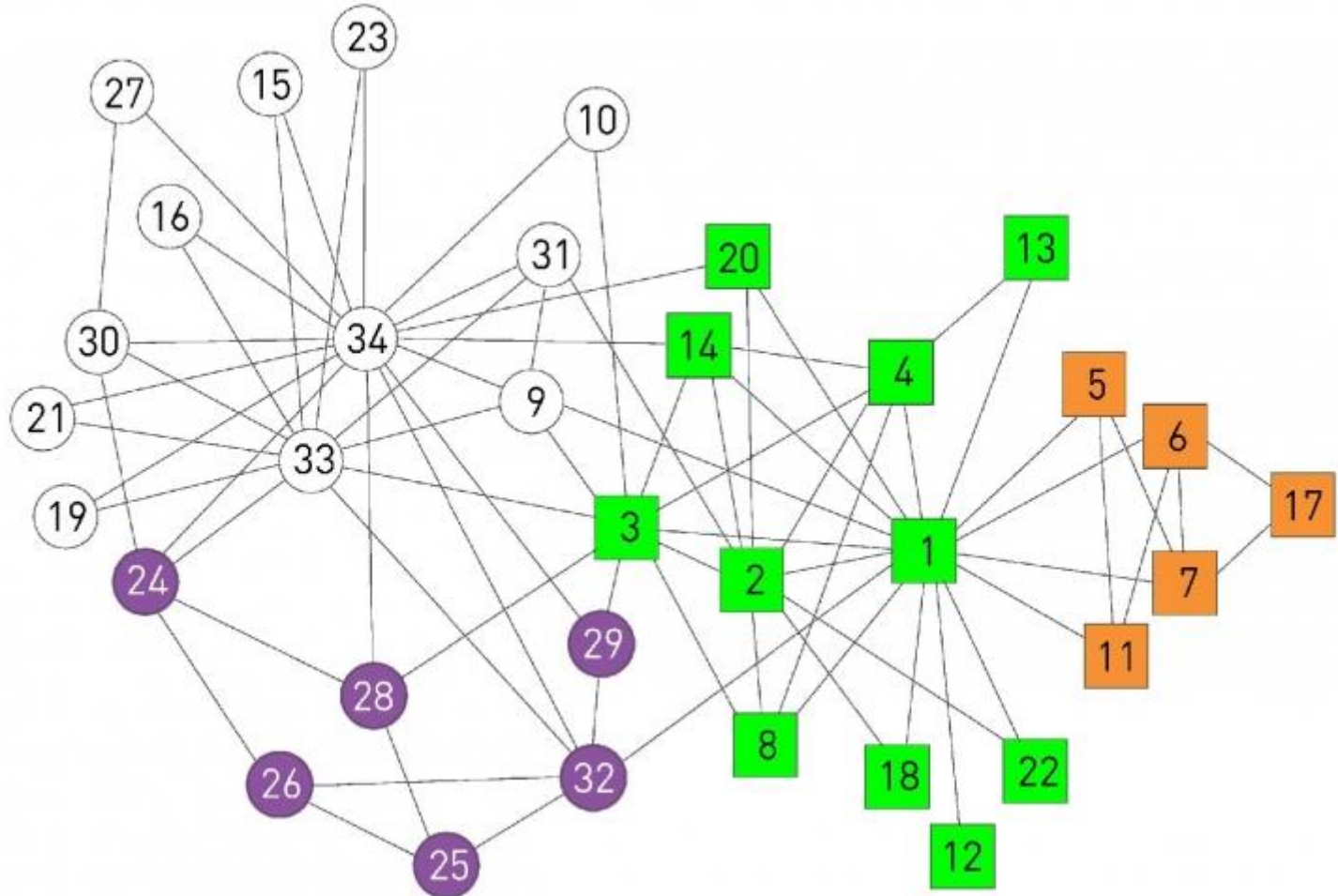
Modularność



Założenie maksymalnej modularności

Dla każdej sieci istnieje podział na grupy, który daje nam najwyższą wartość modularności. Ten podział jest optymalnym podziałem na grupy.

Zachary's Karate Club



Algorytm zachłanny

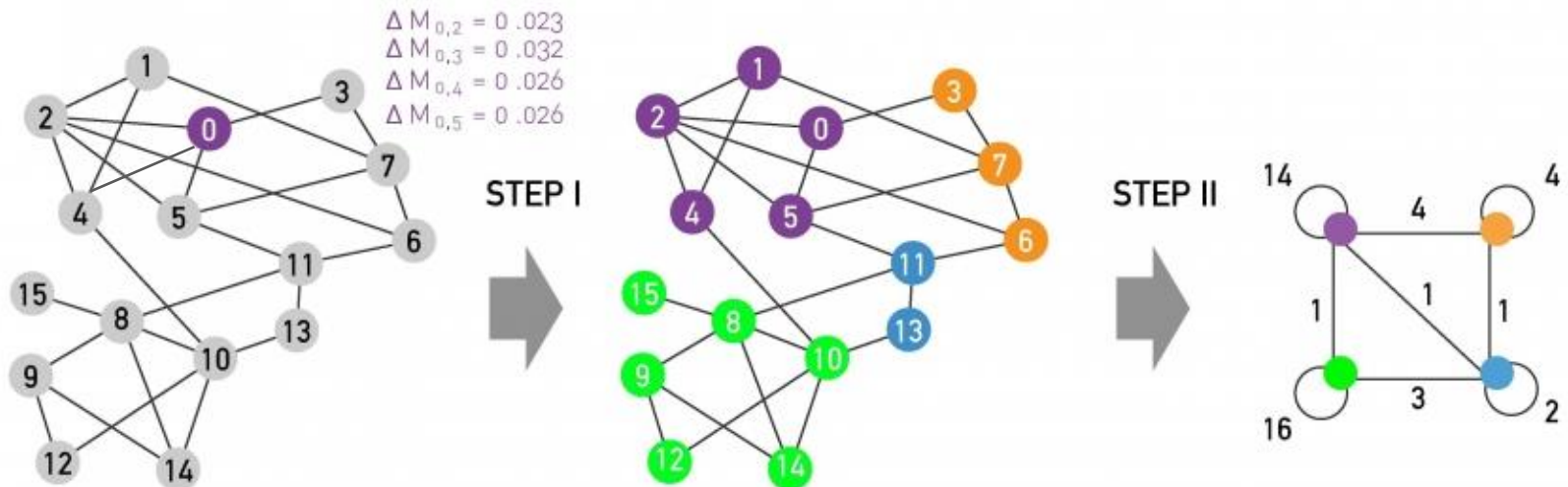
1. Przypisz każdy węzeł do własnej grupy, zaczynając od N grup składających się z pojedynczych węzłów.
2. Sprawdź każdą parę grup połączoną co najmniej jedną krawędzią i oblicz różnicę w modułowości ΔM uzyskaną, jeśli je scalimy. Zidentyfikuj parę społeczności, dla której ΔM jest największe, i połącz je. Zauważ, że modularność jest zawsze obliczana dla całej sieci.
3. Powtarzaj krok 2, aż wszystkie węzły połączą się w jedną grupę, zapisując M dla każdego kroku.
4. Wybierz partycję, dla której wartość M jest największa.

Algorytm zachłanny – Limit rozdzielczości (Resolution Limit)

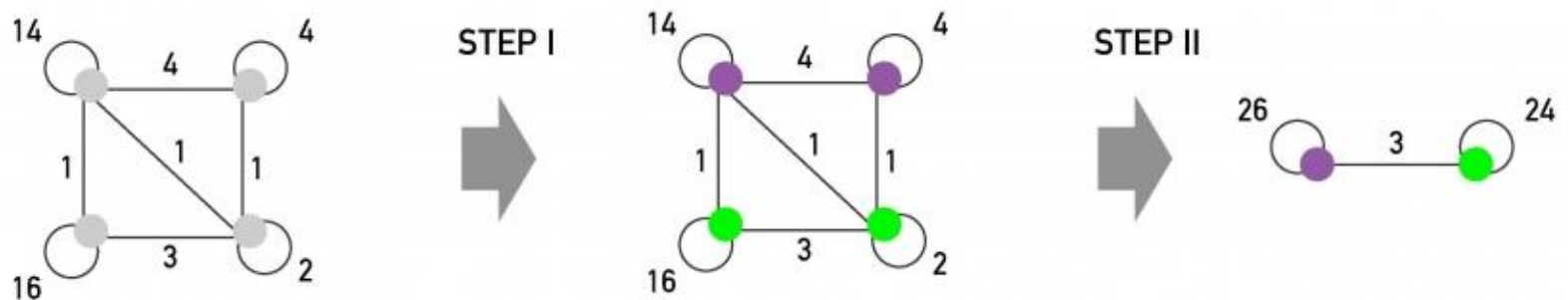
- Jeżeli mamy dwie grupy A i B o sumarycznym stopniu wężła k_A i k_B i jeżeli $k_A + k_B < \sqrt{2L}$ to połączenie tych dwóch grup zawsze da nam przyrost modularności niezależnie od tego czy jest to sensowne połączenie czy nie.
- Maksymalizacja Modularności nie jest w stanie wykrywać grup mniejszych niż limit rozdzielczości co oznacza że np. dla sieci WWW z $L=1\,497\,134$ będzie miało problem z wykryciem grup o $k_C \lesssim 1\,730$.
- Rzeczywiste sieci zawierają wiele małych grup. Biorąc pod uwagę limit rozdzielczości, te małe grupy są łączone w większe grupy, co w konsekwencji prowadzi do błędnego podziału na grupy.

Algorytm Louvain $O(L)$ – Fast modularity optimization

1ST PASS



2ND PASS





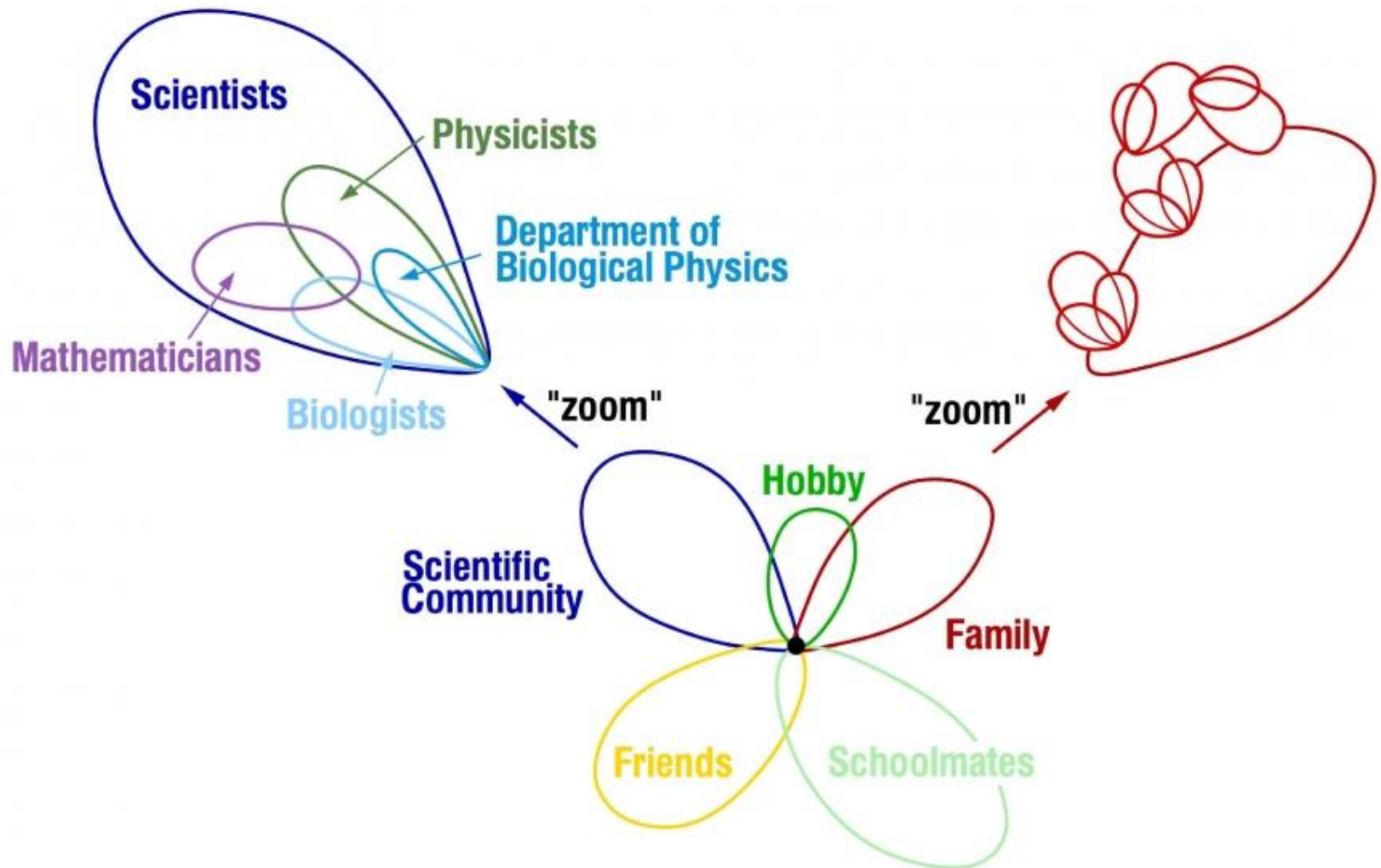
Politechnika
Wrocławska

Grupy nachodzące na siebie



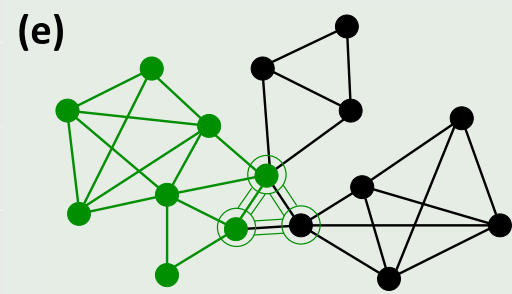
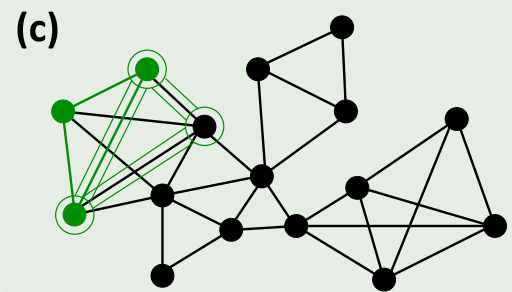
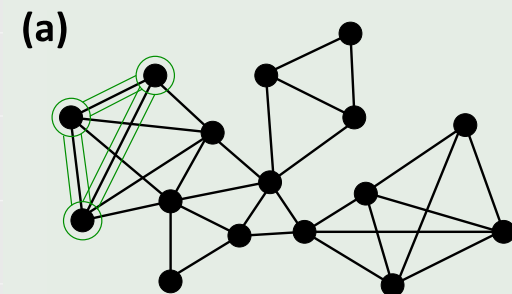
HR EXCELLENCE IN RESEARCH

Grupy nachodzące na siebie

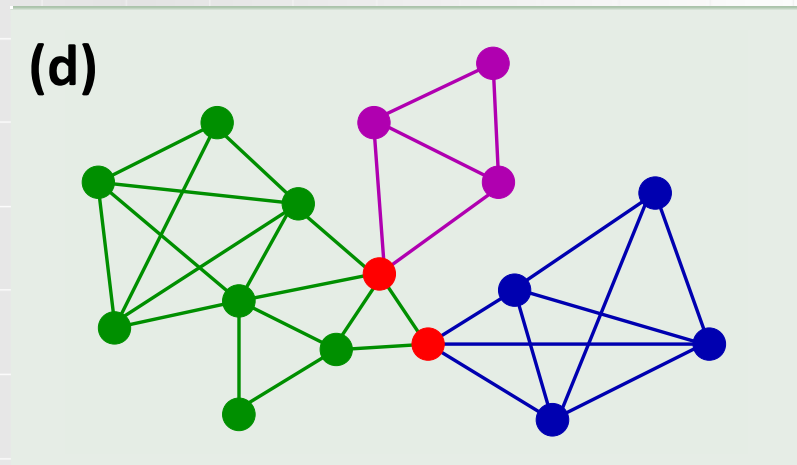
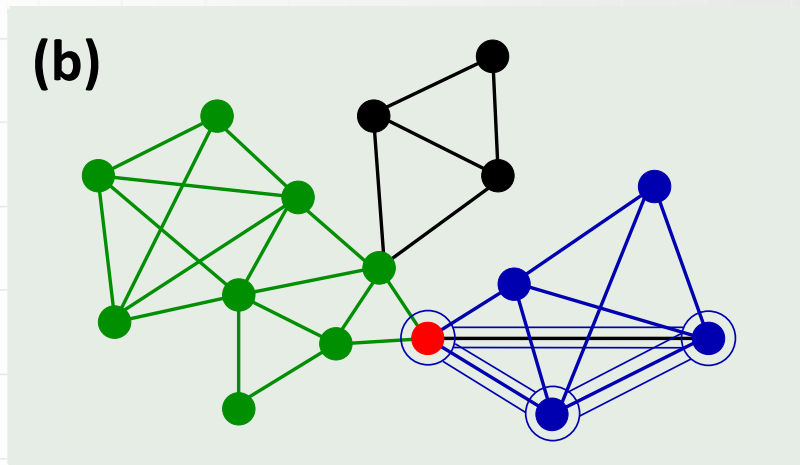


CFinder – algorytm perkolacji klik

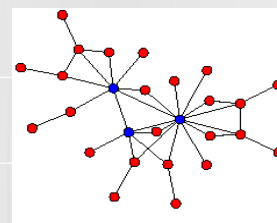
- Dwie k -kliki (grafy pełne składające się z k wierzchołków) są sąsiadujące jeżeli dzielą $k-1$ wierzchołków
- k -grupa to największy połączony podgraf uzyskany przez połączenie sąsiadujących k -klik



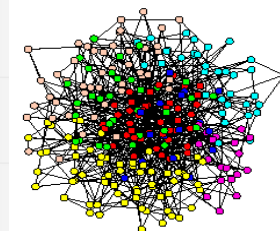
CFinder – algorytm perkolacji klik



- Duża złożoność obliczeniowa
- Nie jesteśmy w stanie znaleźć grup mniejszych niż k
- Liście oraz węzły mające pojedynczych sąsiadów nie należą do żadnej grupy



Scale-free





Politechnika
Wrocławska

Ewaluacja metod wykrywania grup



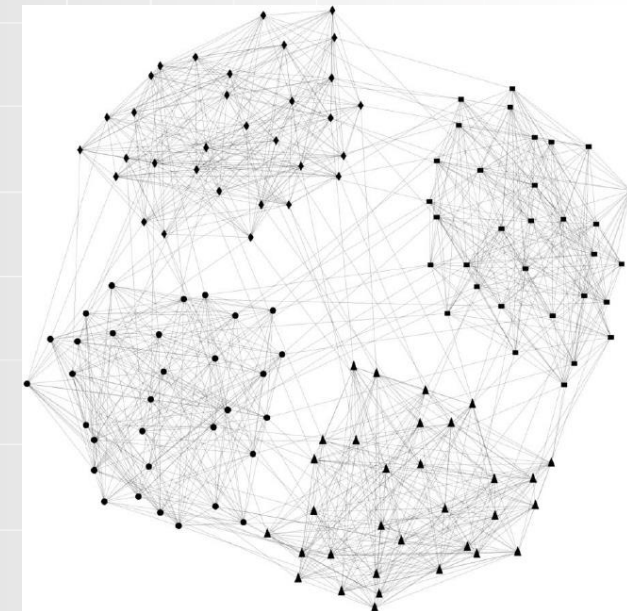
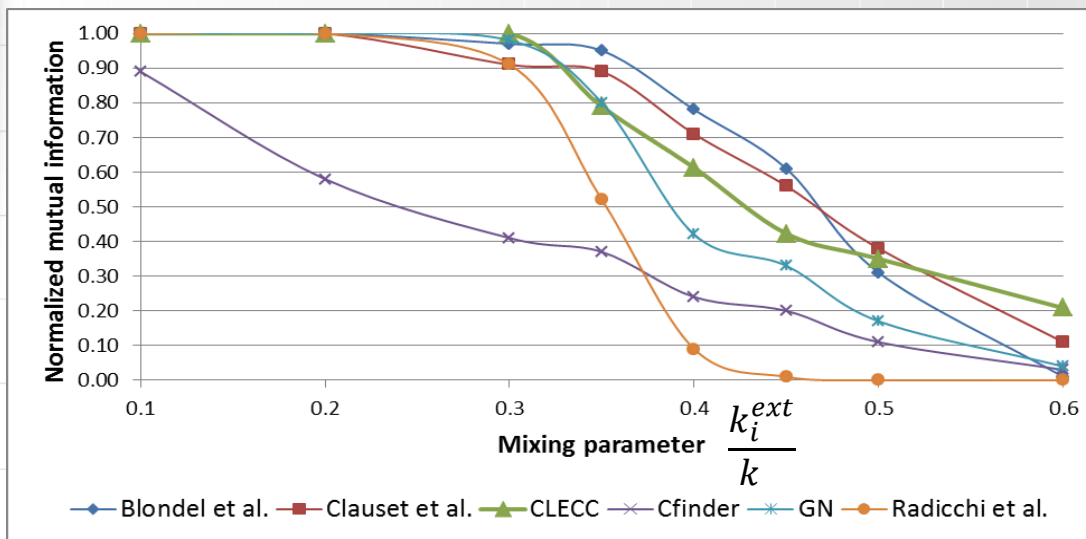
HR EXCELLENCE IN RESEARCH

Ewaluacja metod wykrywania grup

- Sieci referencyjne (Networks with ground-truth communities)
 - Zachary karate club
 - American College football
 - Dolphins
 - <http://snap.stanford.edu/data/index.html#communities>

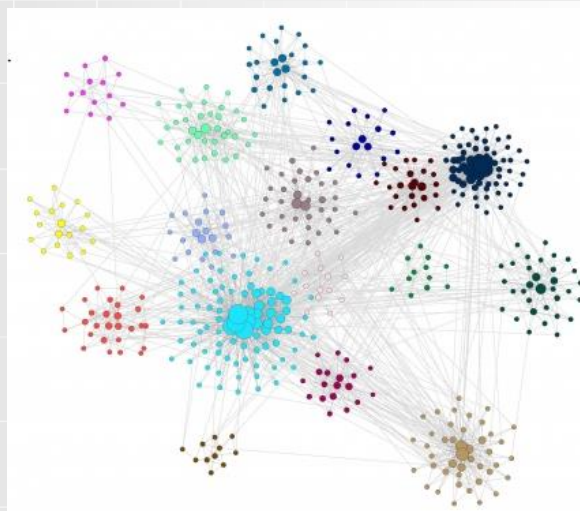
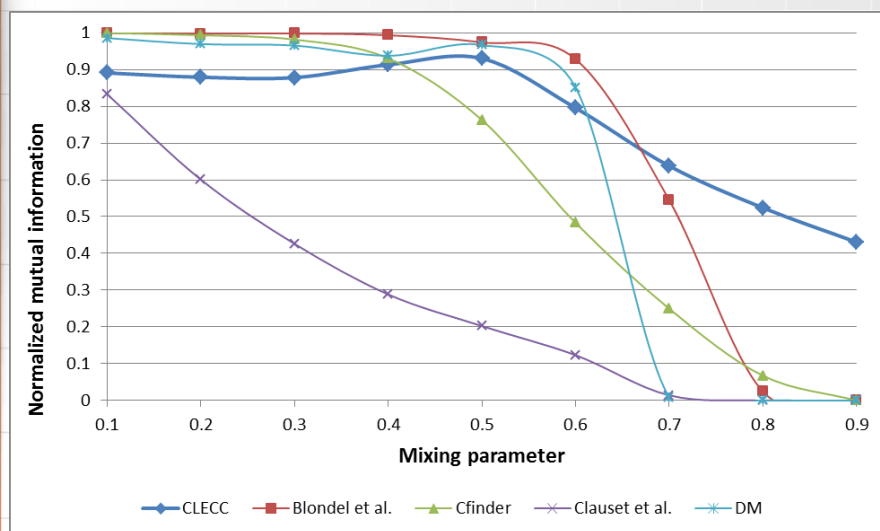
Ewaluacja metod wykrywania grup

- Girvan - Newman Benchmark
 - Sieć ma 128 węzłów podzielonych na 4 grupy po 32 węzły w każdej
 - Każdy węzeł ma stopień $k=16$
 - Losowo łączymy wierzchołki w grupie tak by każdy miał stopień k_i^{int}
 - Losowo łączymy wierzchołki pomiędzy grupami tak by każdy miał stopień $k_i^{ext} = k - k_i^{int}$



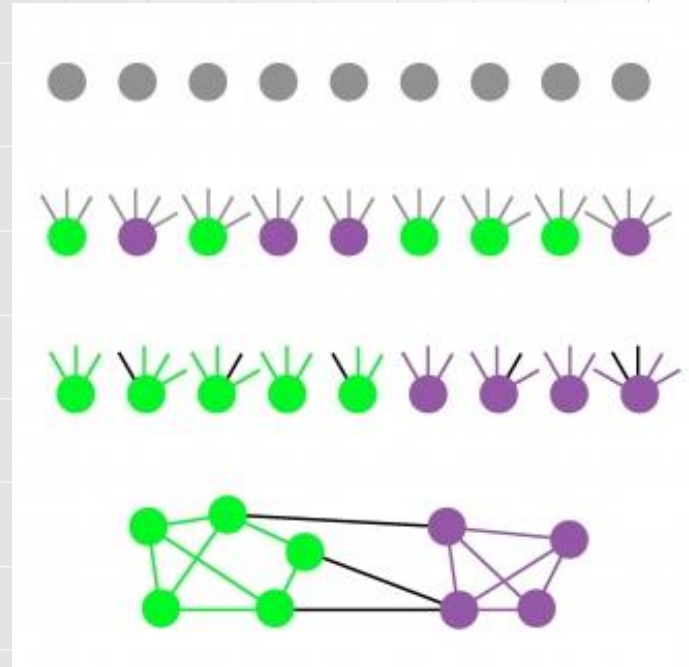
Ewaluacja metod wykrywania grup

- **LFR Benchmark** (L – Lancichinetti, S – Fortunato, R – Radicchi)
 - Bardziej zaawansowany i złożony benchmark
 - Stopień wężła – power law, Wielkość grup – power law
 - możliwość generowania sieci ważonych, skierowanych, z nakładającymi się grupami i sieci wielowarstwowych (mLFR Benchmark¹)



LFR Benchmark

- Rozpocznij od N izolowanych wierzchołków.
- Przypisz każdy wierzchołek do grupy wielkości N_c gdzie rozkład N_c jest zgodny z rozkładem power-law
- Przypisz każdemu wierzchołkowi i stopień wężła k_i , rozkład stopni wężła powinien być zgodny z rozkładem power-law
- Każdy wierzchołek i otrzymuje wewnętrzny stopień $(1-\mu)k_i$ zewnętrzny stopień μk_i .
- Wszystkie wierzchołki w grupie są losowo łączone by spełniony był ich stopień wewnętrzny, następnie losowo łączymy wierzchołki pomiędzy grupami tak by spełniony był ich stopień zewnętrzny.



Ewaluacja metod wykrywania grup

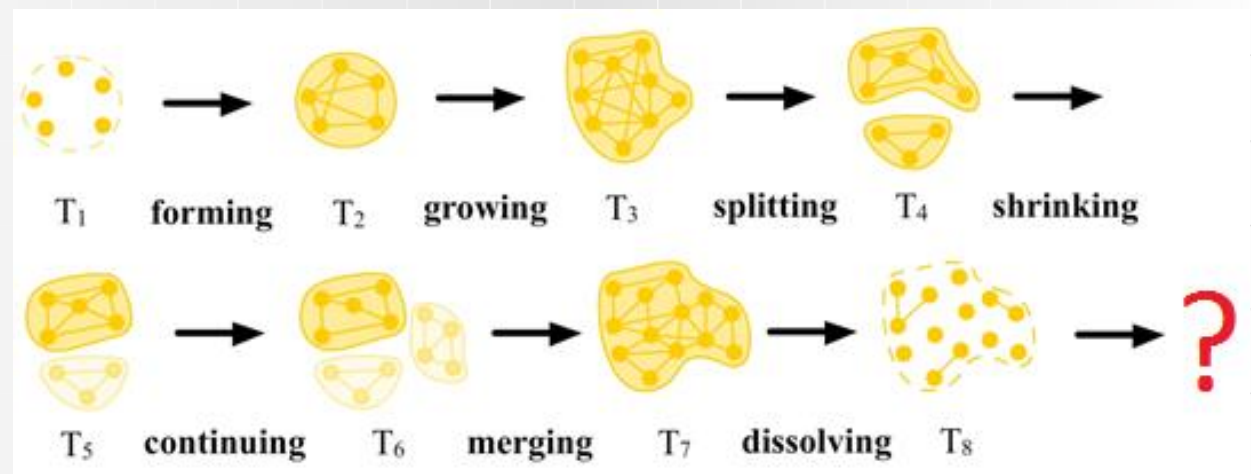
Modelowe grupy $A = \{G_1, G_2, \dots, G_a\}$,

Wykryte grupy $B = \{G_1, G_2, \dots, G_b\}$.

- Indeks Jaccarda (Jaccard index) $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$
- (Znormalizowana) Informacja wzajemna ((Normalized) mutual information)
$$NMI(A, B) = \frac{-2 \sum_{i=1}^a \sum_{j=1}^b n_{ij} \log \left(\frac{n_{ij} n}{n_i n_j} \right)}{\sum_{i=1}^{c_A} n_i \log \left(\frac{n_i}{n} \right) + \sum_{j=1}^{c_B} n_j \log \left(\frac{n_j}{n} \right)}$$
- Zmienność informacji (variation of information/shared information distance)
- Modularność
- Itd.

Co dalej?

- Wykrywanie grup jest ciekawe...
- ... ale przewidywanie co się stanie z grupą w przyszłości jest jeszcze ciekawsze.





Politechnika
Wrocławska

GED - Group Evolution Detection



HR EXCELLENCE IN RESEARCH

Bródka P., Saganowski S., Kazienko P.: **GED: The Method for Group Evolution Discovery in Social Networks**. Social Network Analysis and Mining, 3(1), 2013, pp. 1-14

Temporalna sieć społeczna

Strumień
krawędzi

...

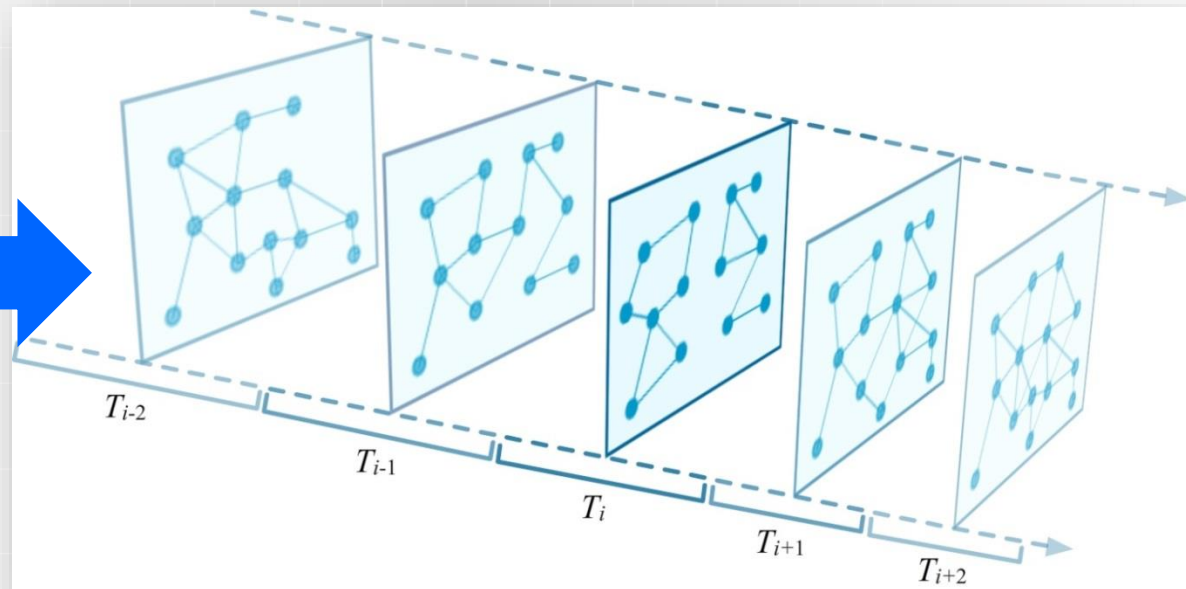
$\langle v_{10}, v_{20}, t_6 \rangle$

...

$\langle v_i, v_j, t_k \rangle$

...

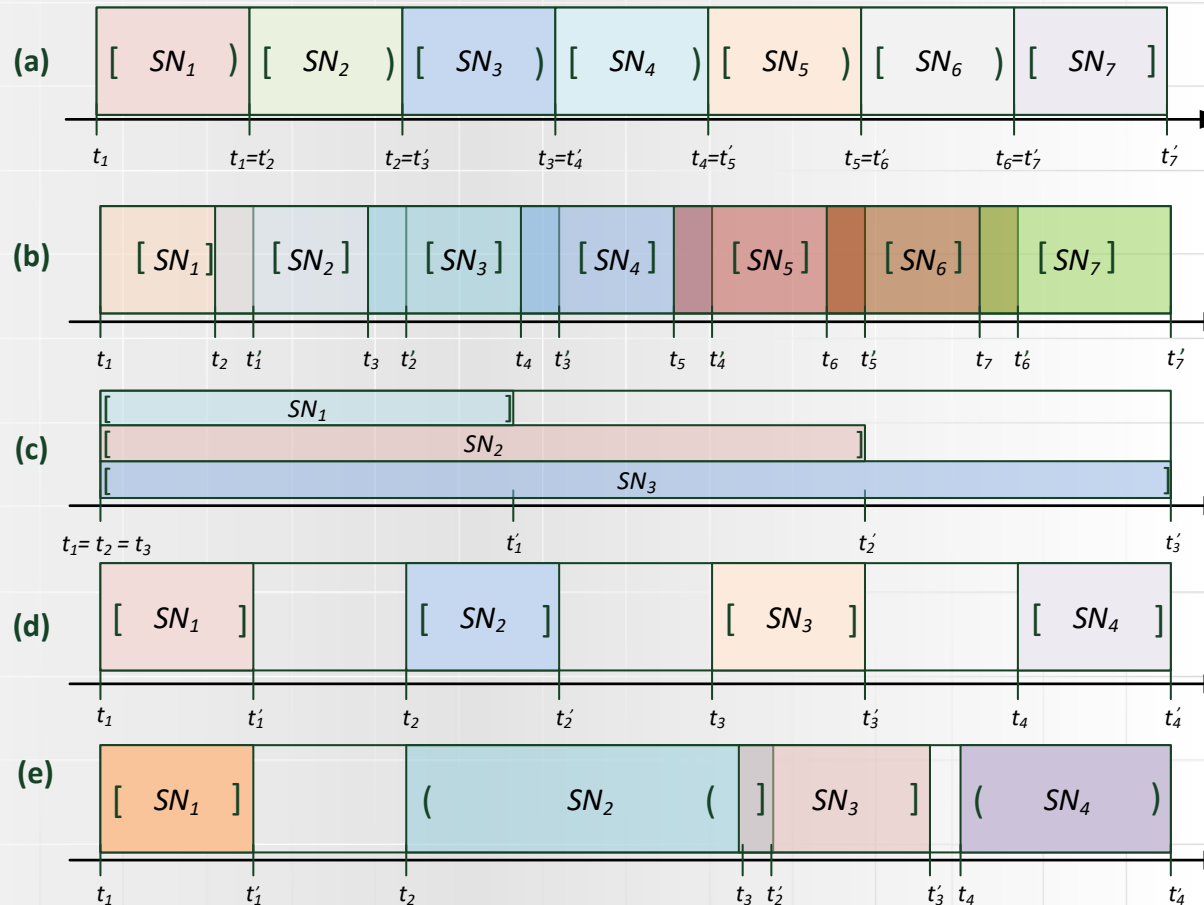
np. maile



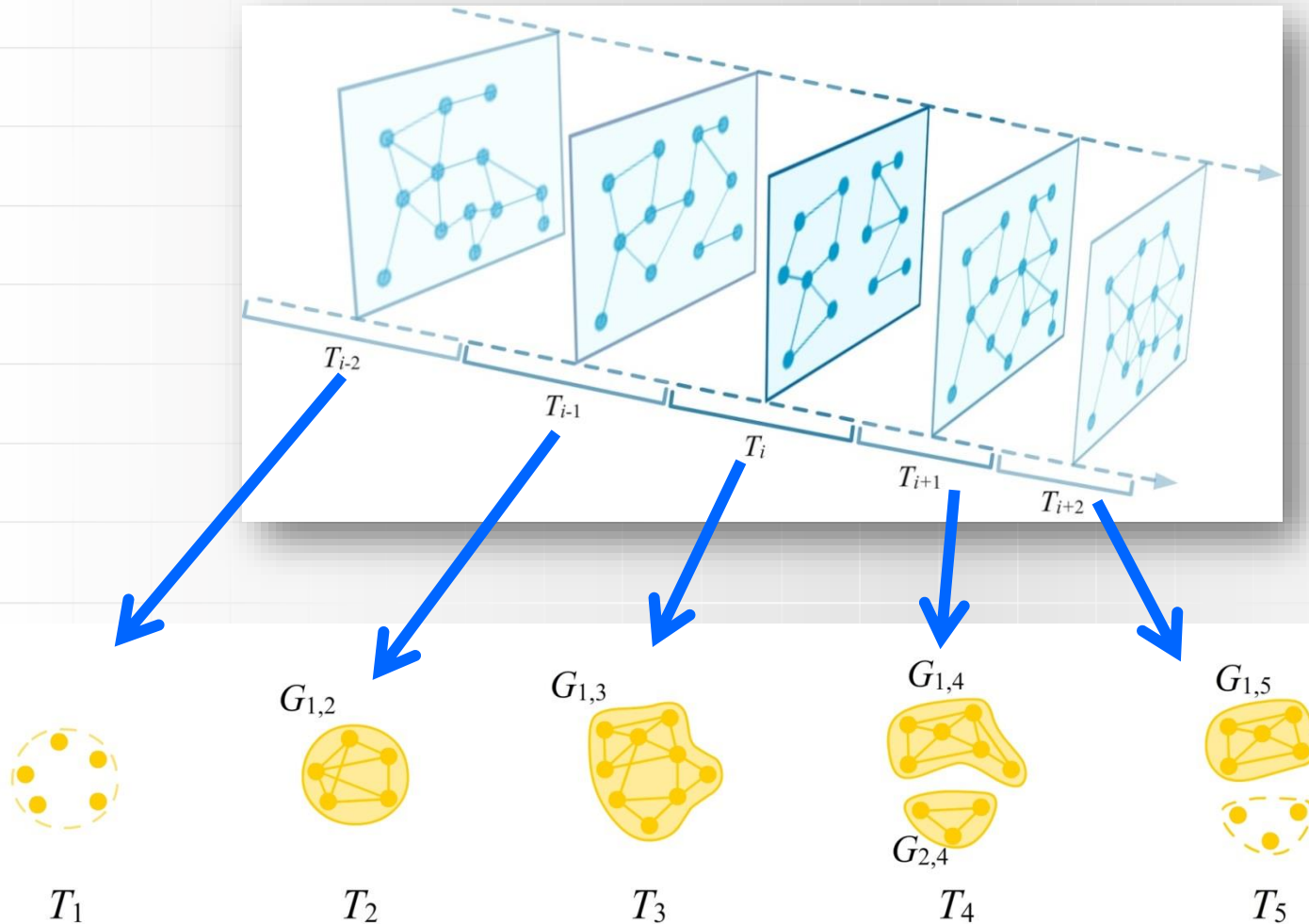
Temporalna sieć społeczna
Temporal Social Network TSN

Sieć temporalna TSN

- a) Okna sąsiadujące
- b) Okna nachodzące
- c) Okna rozszerzające się
- d) Okna sąsiadujące z przerwą
- e) Mieszane



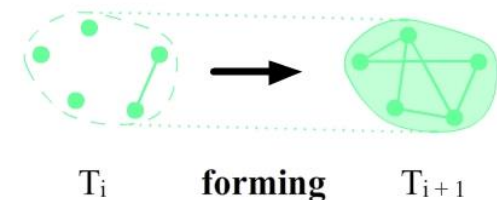
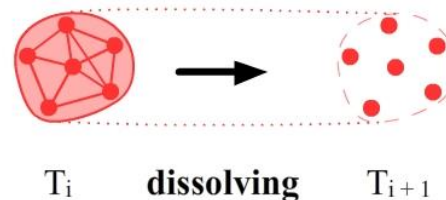
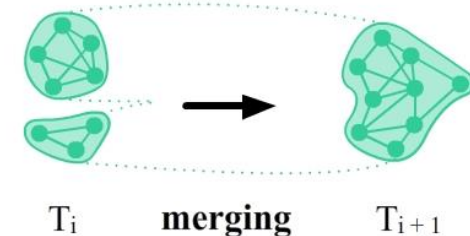
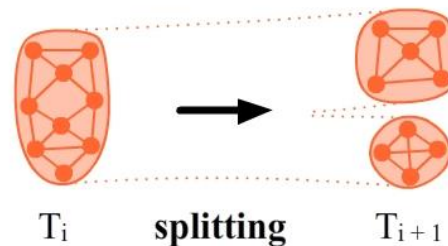
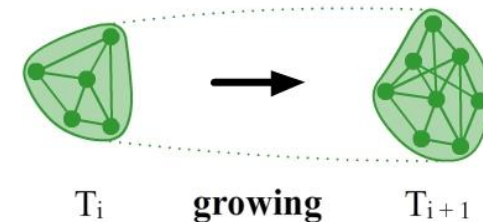
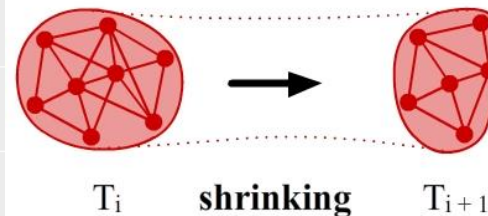
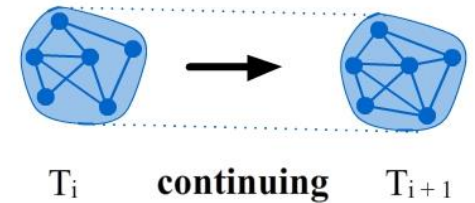
GED - Group Evolution Detection



GED - Group Evolution Detection

- Ewolucja grupy to sekwencja następujących po sobie zdarzeń w kolejnych oknach czasowych w ramach TSN

- Kontynuacja** Continuing
- Kurczenie** Shrinking
- Wzrost** Growing
- Podział** Splitting
- Łączenie** Merging
- Rozpad** Dissolving
- Formacja** Forming



GED

$$I(G_1, G_2) = \frac{\overbrace{|G_1 \cap G_2|}^{\text{group quantity}}}{|G_1|} \cdot \frac{\sum_{x \in (G_1 \cap G_2)} SP_{G_1}(x)}{\underbrace{\sum_{x \in (G_1)} SP_{G_1}(x)}_{\text{group quality}}}$$

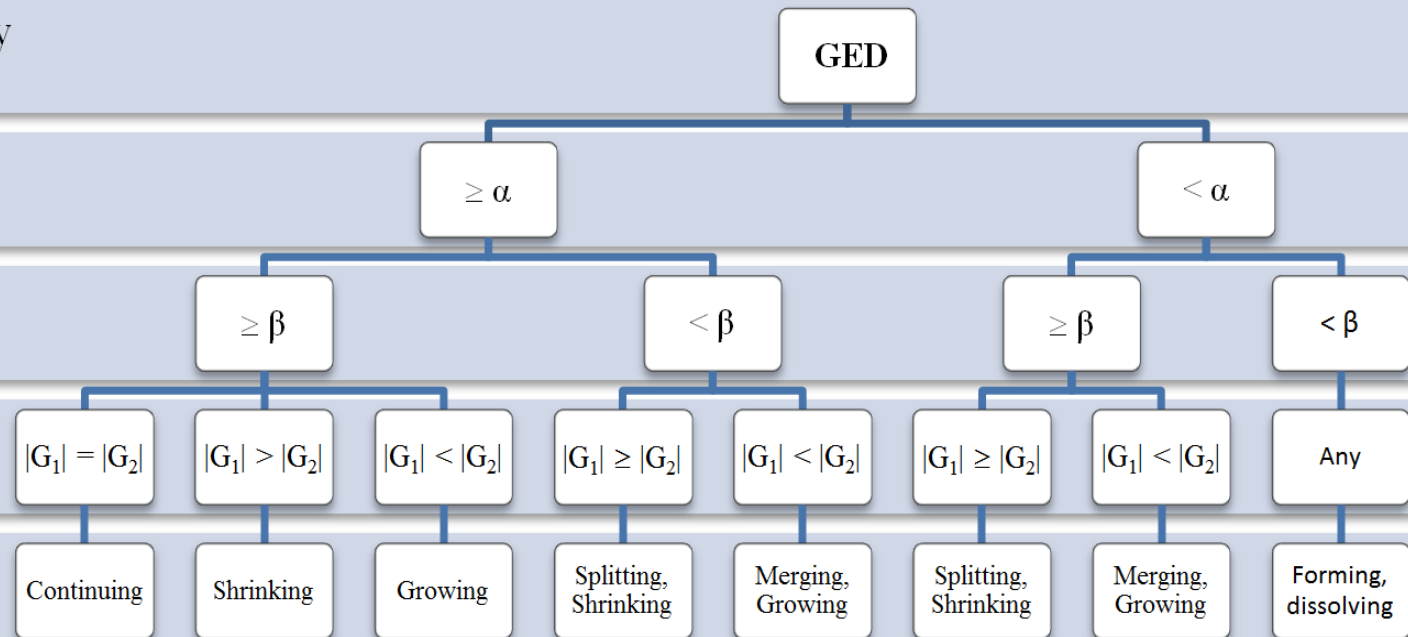
Group Evolution Discovery in Social Networks

Calculating $I(G_1, G_2)$

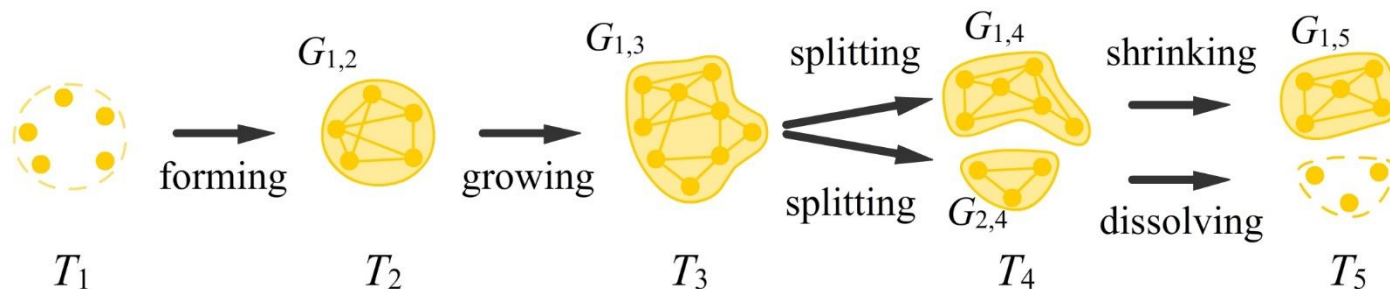
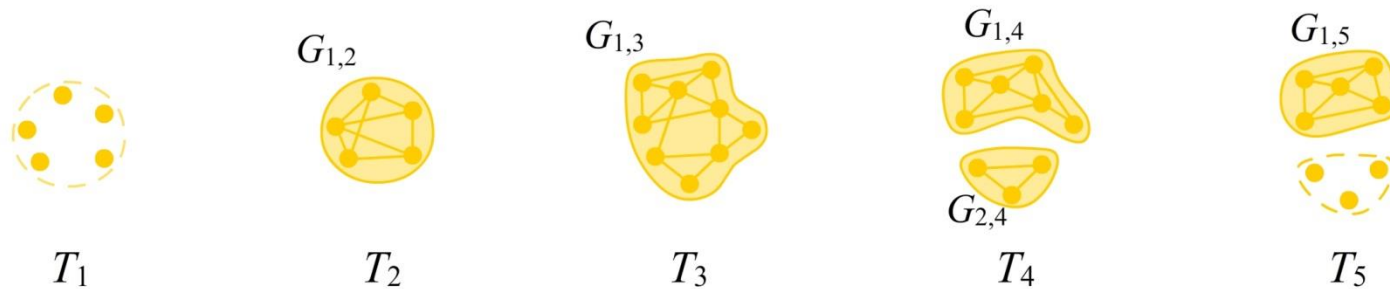
Calculating $I(G_2, G_1)$

Comparing groups sizes

Assigning the event



GED - Group Evolution Detection





Politechnika
Wrocławska

GEP - Group Evolution Prediction



HR EXCELLENCE IN RESEARCH

Saganowski S., Bródka P., Koziarski M., Kazienko P. (2019) . **Analysis of group evolution prediction in complex networks.** PLoS ONE, 14(10): e0224194

Temporalna sieć społeczna

Strumień
krawędzi

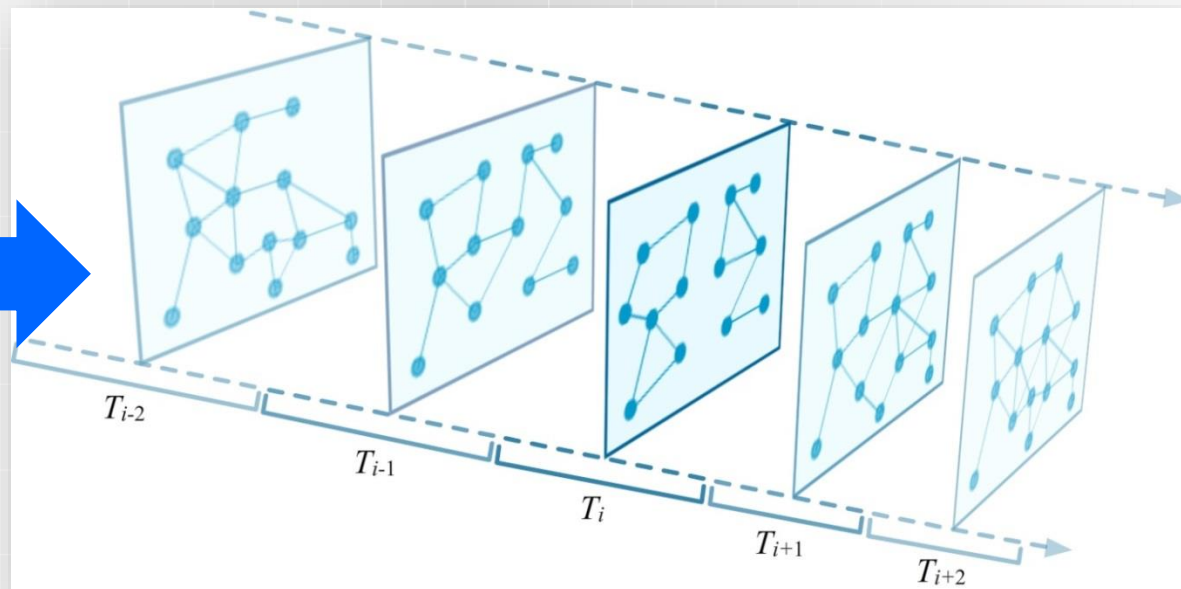
...

$\langle v_{10}, v_{20}, t_6 \rangle$

...

$\langle v_i, v_j, t_k \rangle$

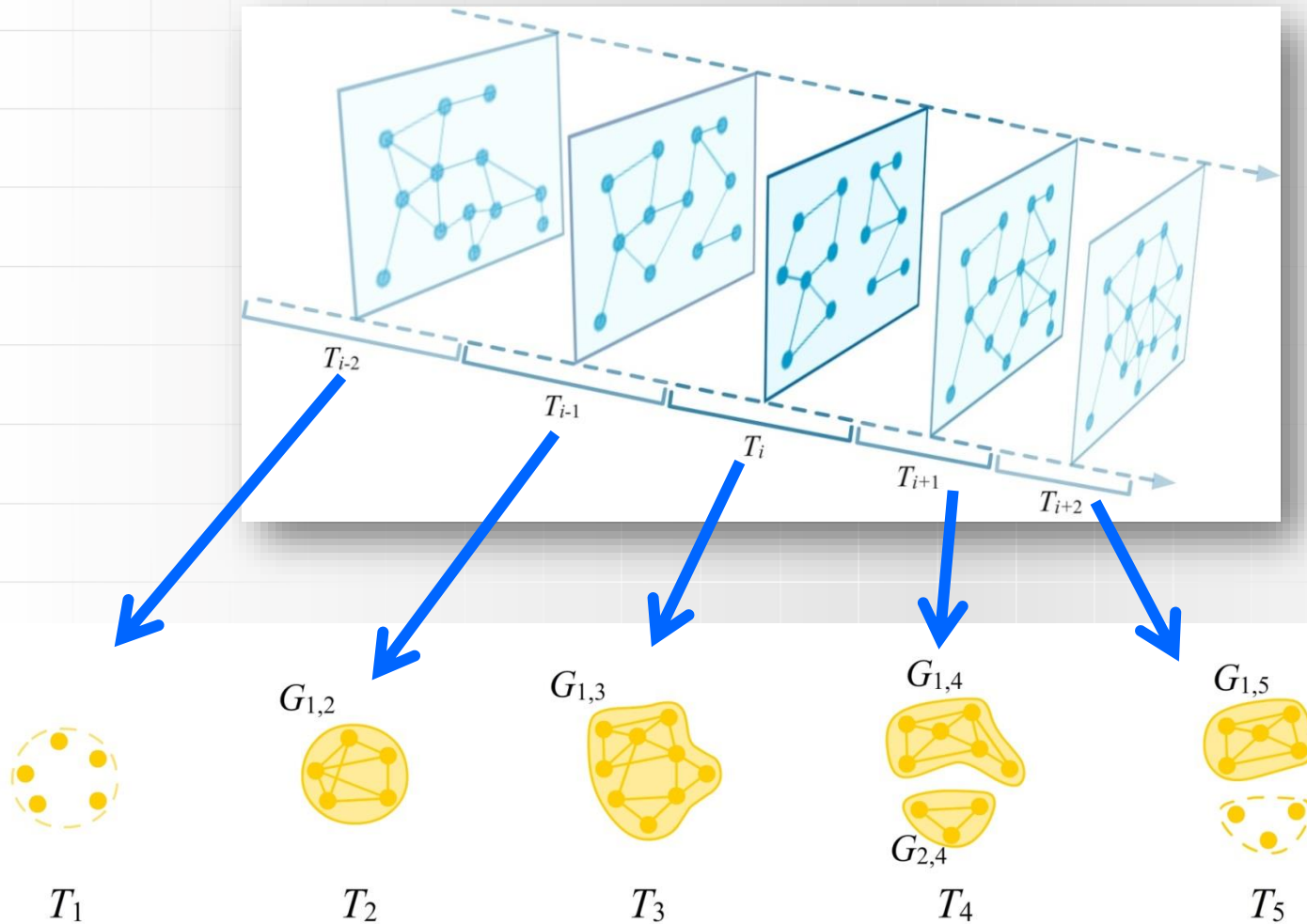
...



np. maile

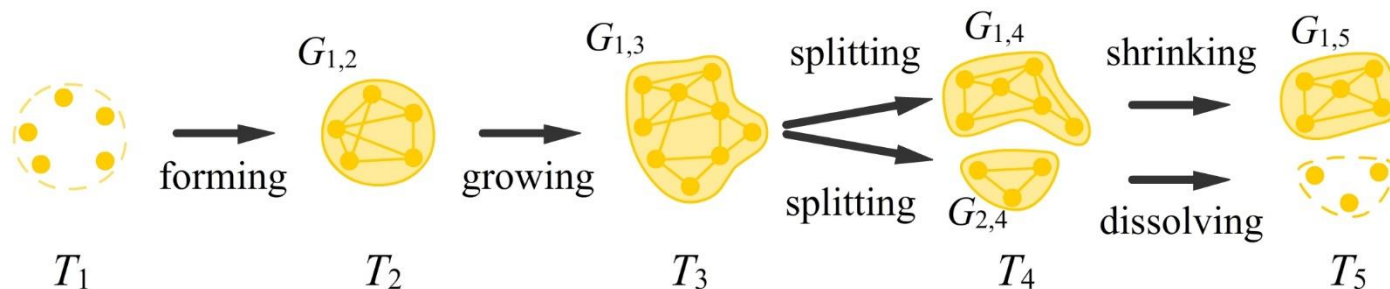
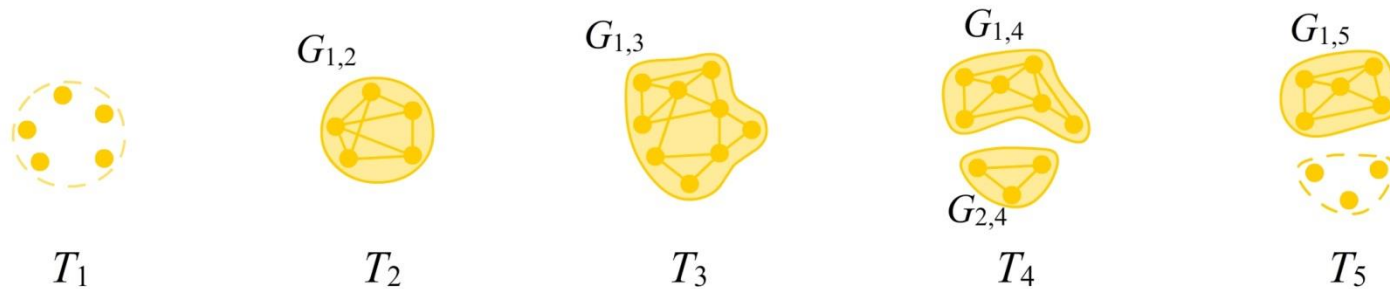
Temporalna sieć społeczna
Temporal Social Network TSN

GEP - Group Evolution Prediction



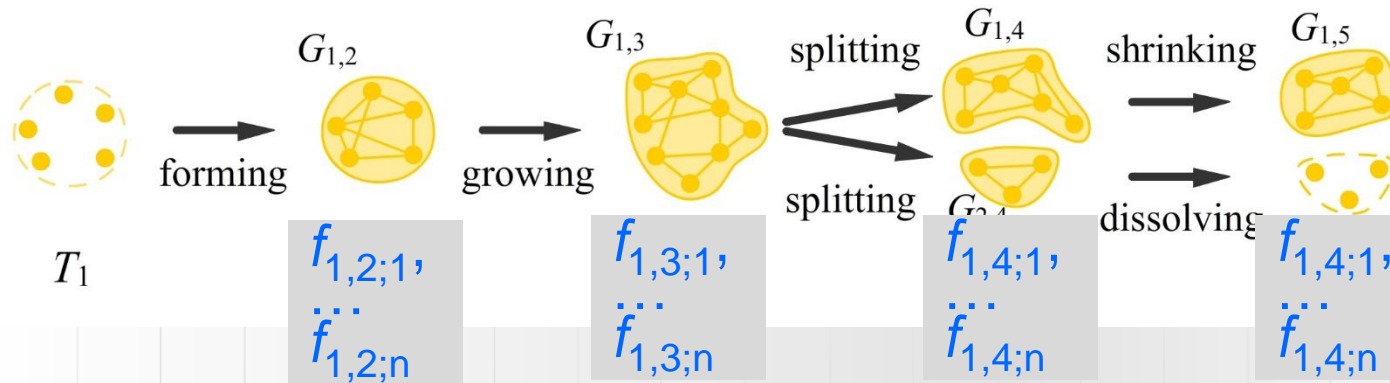
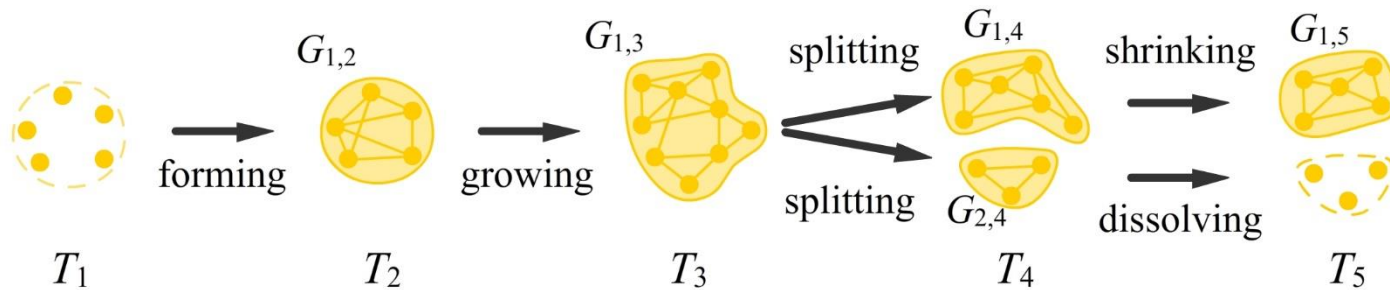
$$IS \xrightarrow{TWT} TW \xrightarrow{NT} TSN \xrightarrow{CDM} G$$

GEP - Group Evolution Prediction



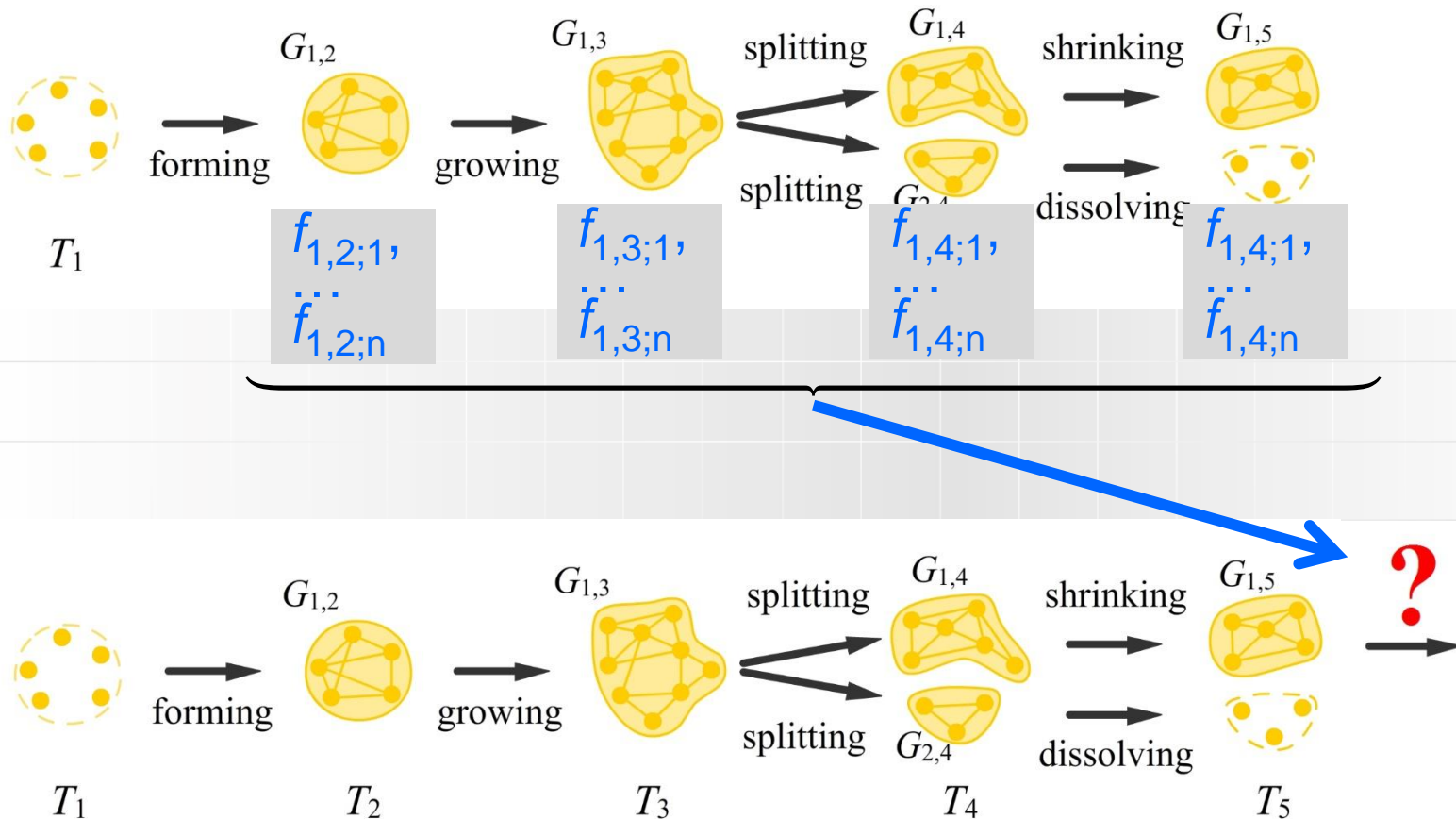
$$IS \xrightarrow{TWT} TW \xrightarrow{NT} TSN \xrightarrow{CDM} G \xrightarrow{CETM} EC$$

GEP - Group Evolution Prediction



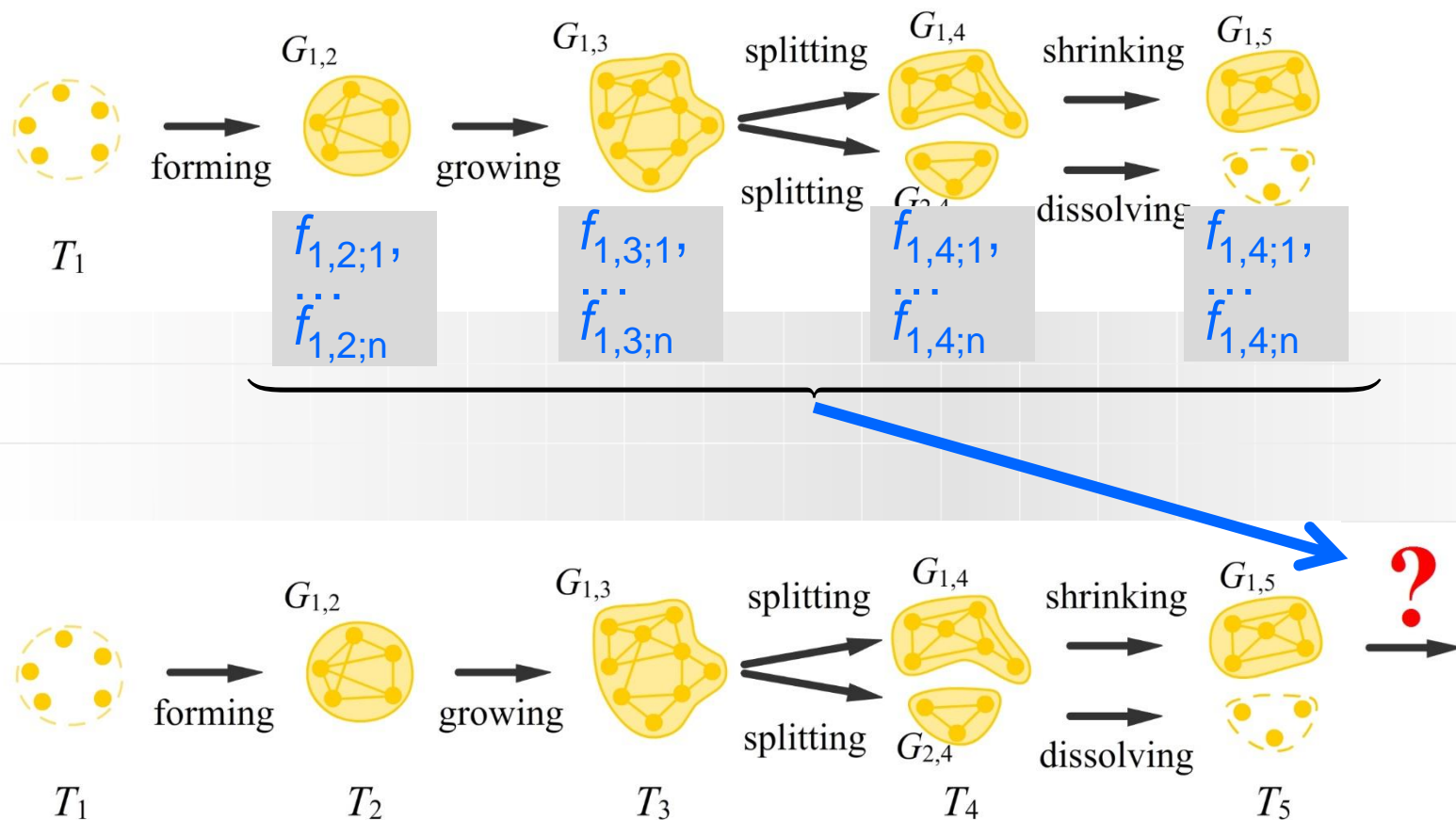
$$IS \xrightarrow{TWT} TW \xrightarrow{NT} TSN \xrightarrow{CDM} G \xrightarrow{CETM} EC \xrightarrow{FE} F$$

GEP - Group Evolution Prediction



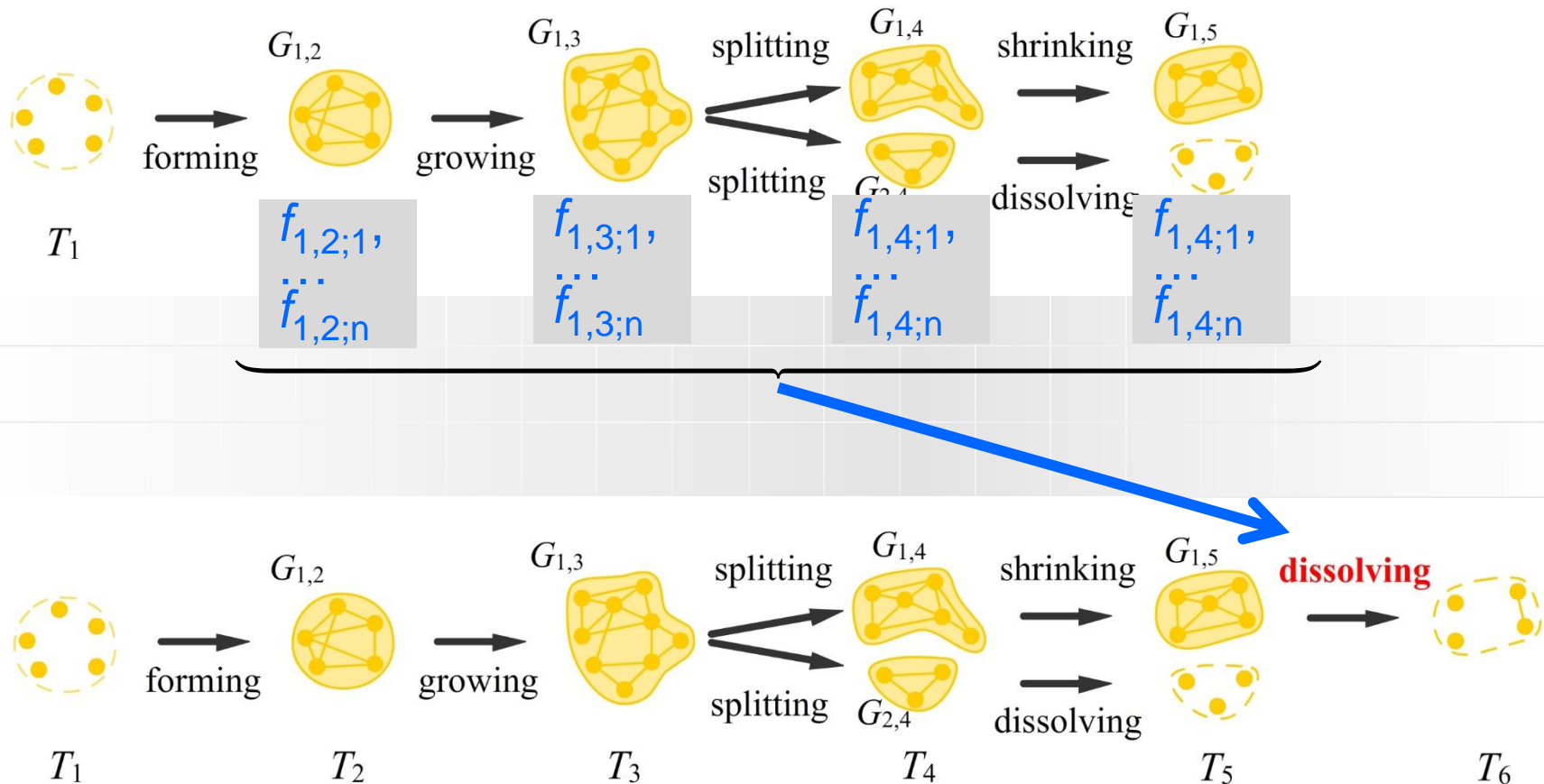
$$IS \xrightarrow{TWT} TW \xrightarrow{NT} TSN \xrightarrow{CDM} G \xrightarrow{CETM} EC \xrightarrow{FE} F$$

GEP method



$$IS \xrightarrow{TWT} TW \xrightarrow{NT} TSN \xrightarrow{CDM} G \xrightarrow{CETM} EC \xrightarrow{FE} F$$

GEP - Group Evolution Prediction



$$IS \xrightarrow{TWT} TW \xrightarrow{NT} TSN \xrightarrow{CDM} G \xrightarrow{CETM} EC \xrightarrow{FE} F \xrightarrow{classif.} CH$$

Pytania

