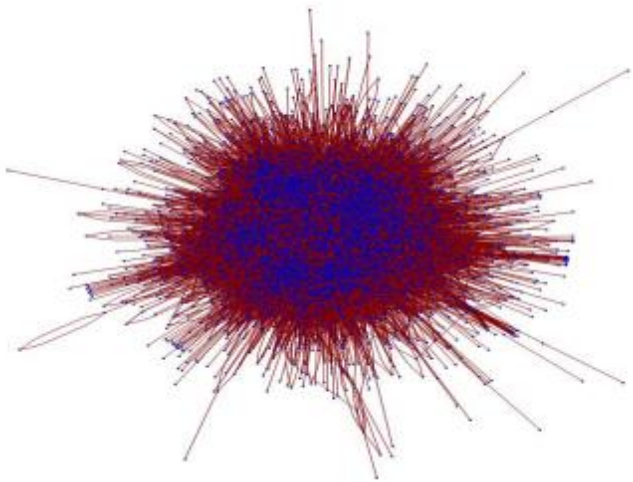
Próbkowanie sieci złożonych

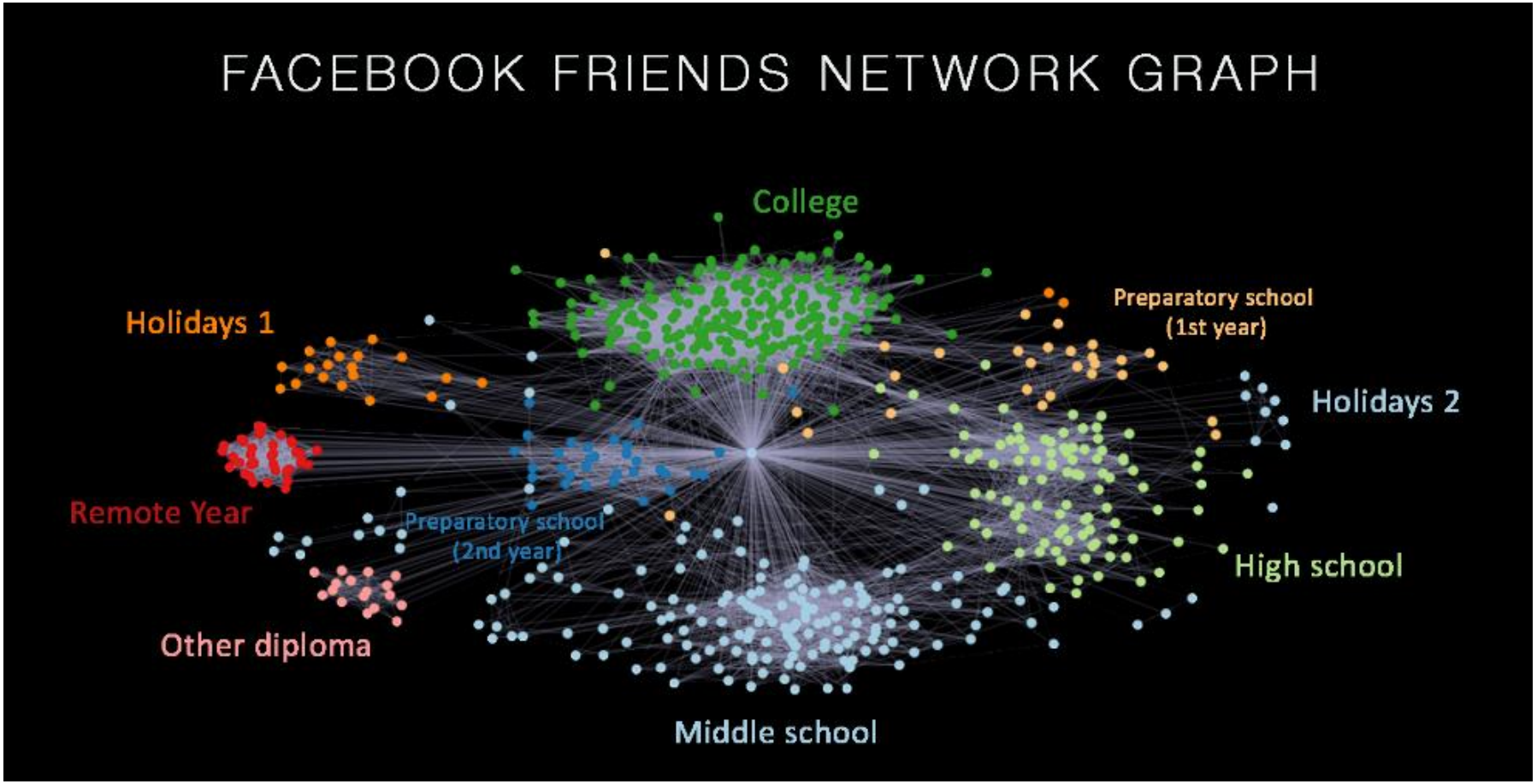
Jarosław Jankowski

Wprowadzenie

- Nauka o sieciach i systemy złożone
- Obliczeniowa nauka o sieciach
  - Dlaczego nabrała znaczenia?
  - Jakie stawiane są cele?
  - Jakie nowe wyzwania?
- Analiza sieci złożonych
- Typy sieci
- Jakie korzyści?



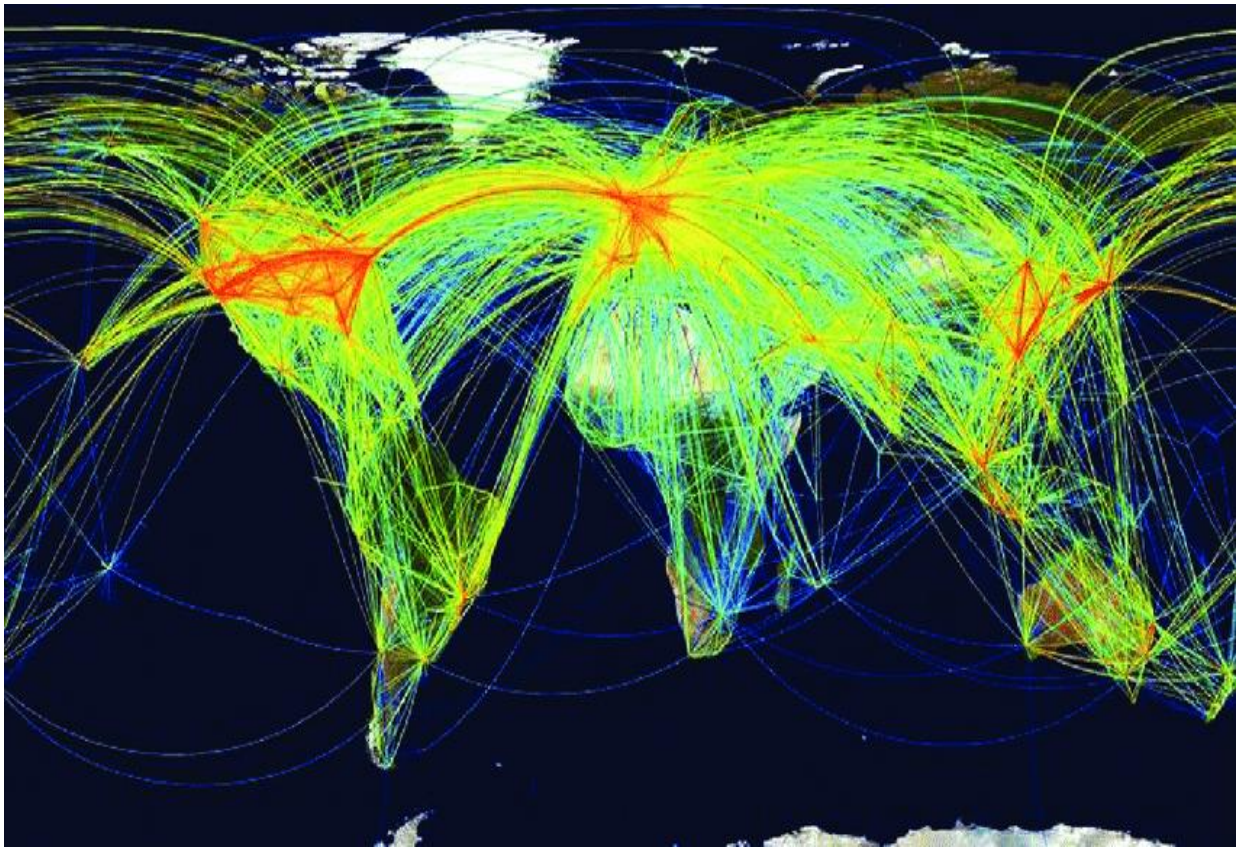
Przykład (1):
Media społecznościowe



https://i.redd.it/p65ghpqrfot01.png

:: Powiązania społeczne :: Dyfuzja informacji

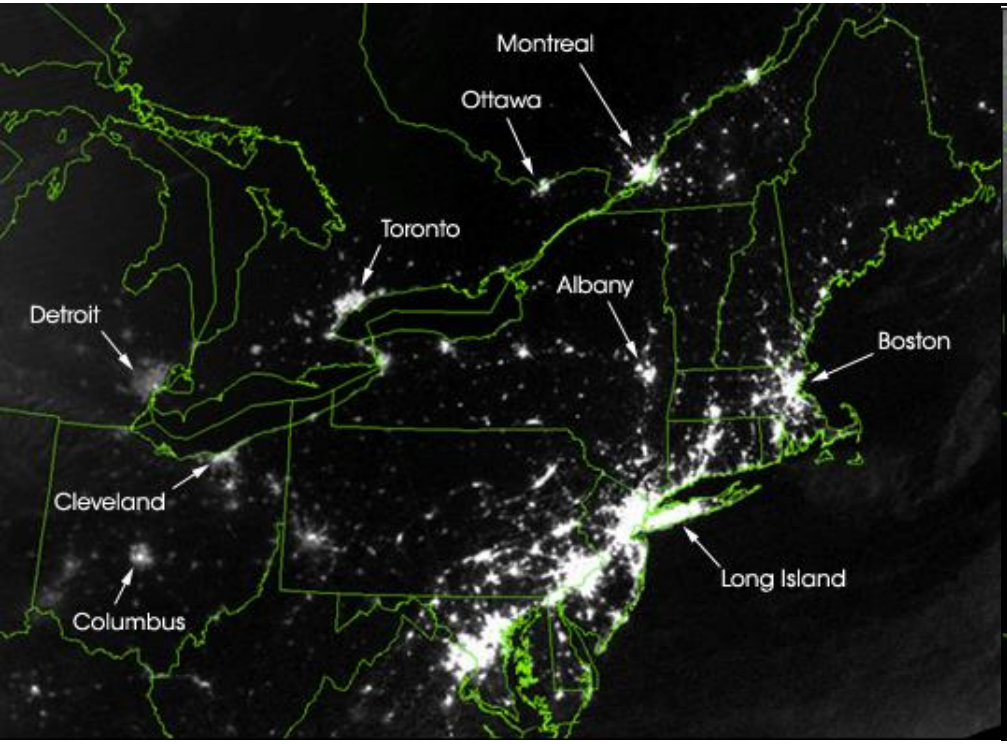
Przykład (2): Epidemiologia

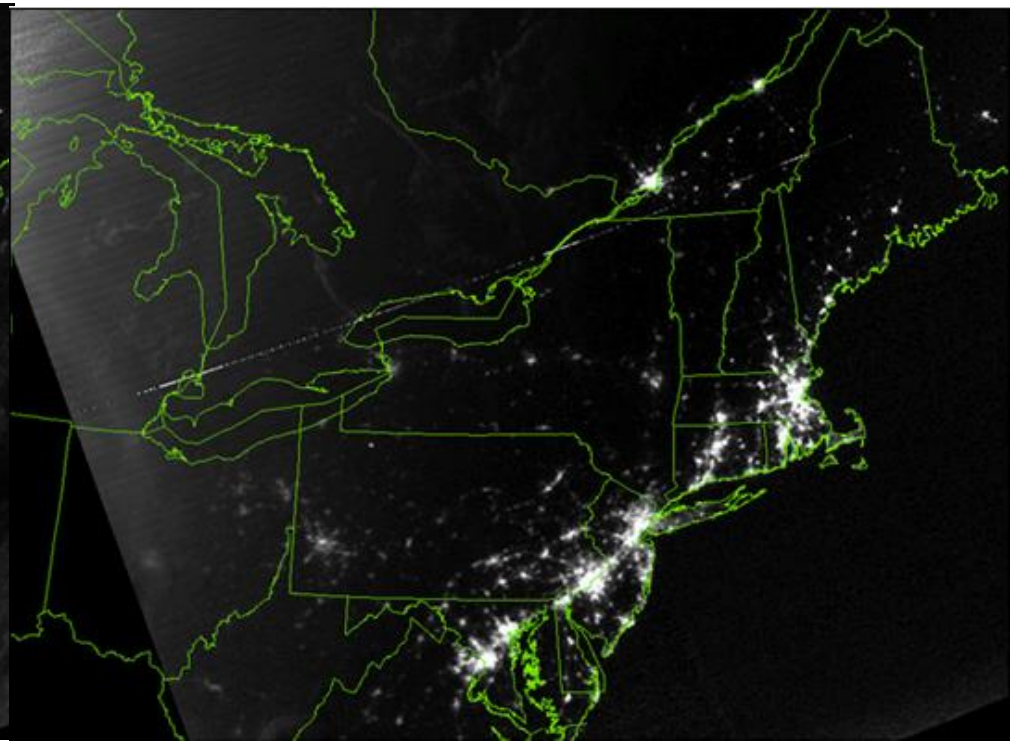


- Identyfikacja źródeł infekcji
- Monitorowanie dynamiki procesów
- **Ograniczanie zasięgu**

https://www.researchgate.net/publication/322204357_Strengthening_Post_Ebola_Health_Systems_From_Response_to_Resilience_in_Guinea_Liberia_and_Sierra_Leone/figures?lo=1

Przykład (3):
15 sierpnia 2003 Blackout.

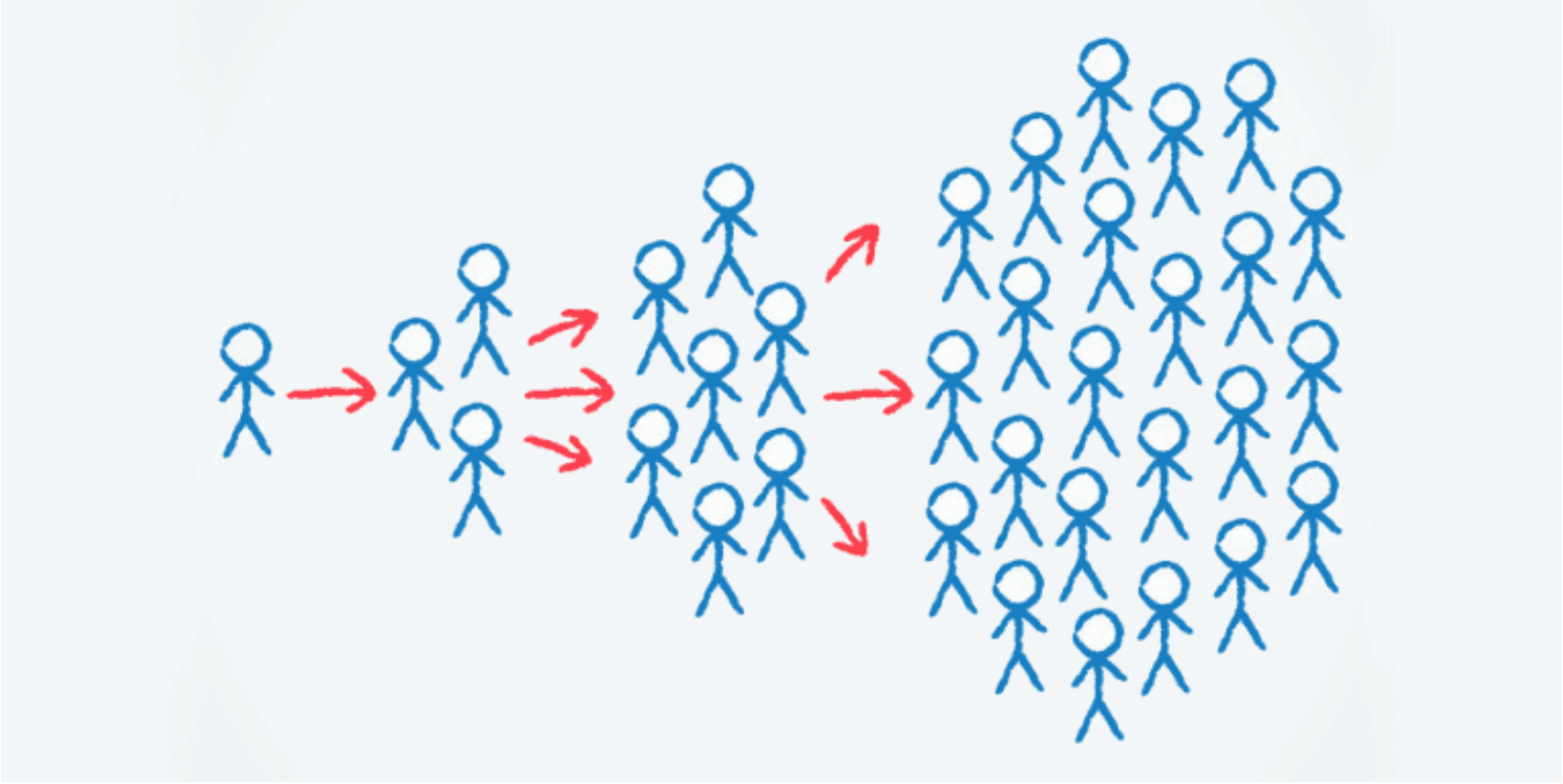




Sierpień 14, 2003:9: 29pm
EDT
20 godzin przed

Sierpień 15, 2003:9: 14pm
EDT
7 godzin po

Przykład (4):
Marketing wirusowy



https://joshfechter.com/wp-content/uploads/2019/03/viral-marketing-definition.png

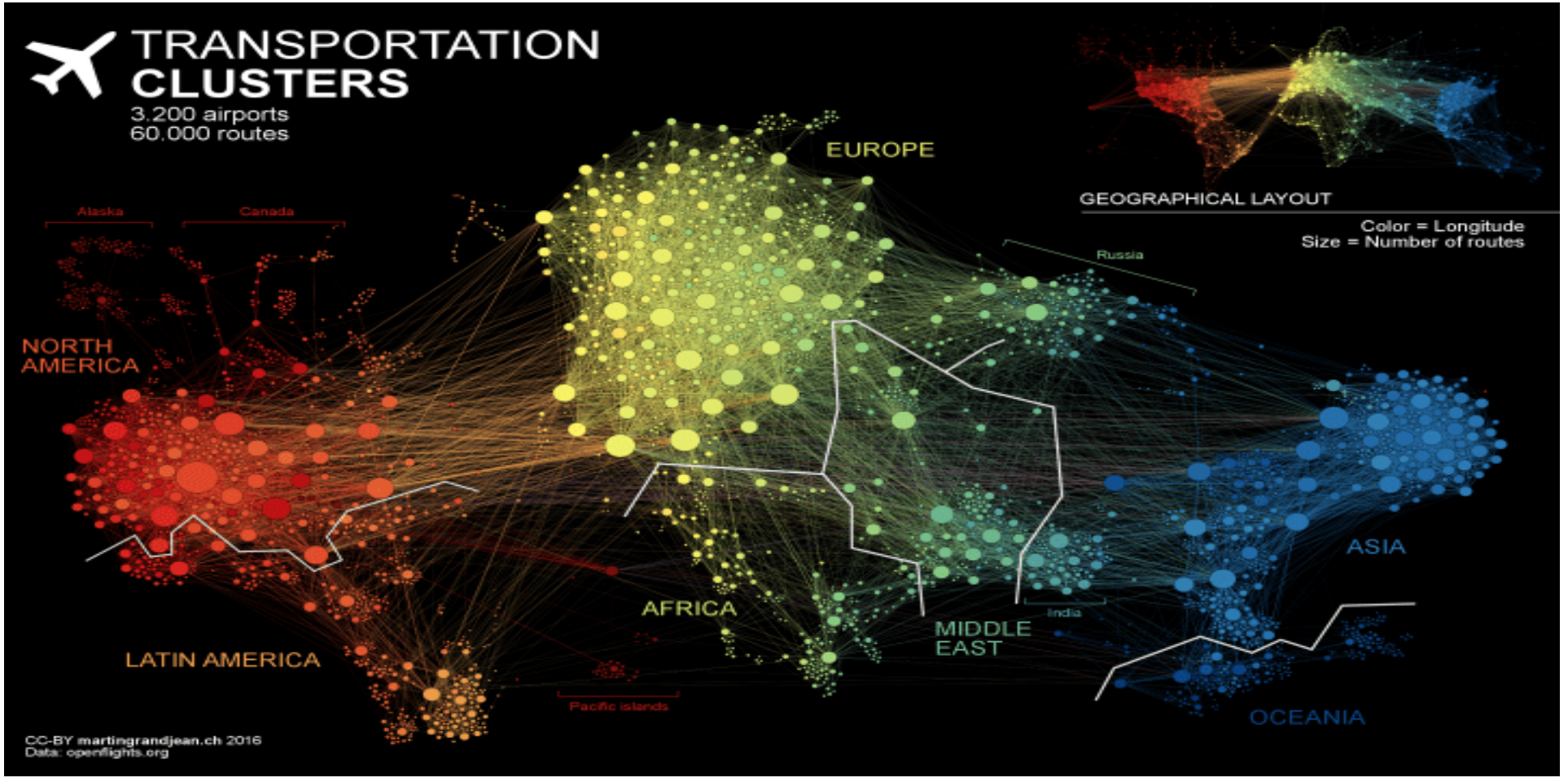
- Rozpoczynanie akcji marketingowych
- **Monitorowanie skuteczności**

Przykład (5): Fake news

- Klasyfikacja źródeł informacji
- Analizy wpływu
- Przeciwdziałanie

https://science.sciencemag.org/content/359/6380/1146

Przykład (6): Sieci transportowe



- Analiza sieci połączeń

- https://i0.wp.com/flowingdata.com/wp-content/uploads/2016/05/airports-world-network.png?fit=720%2C480&ssl=1

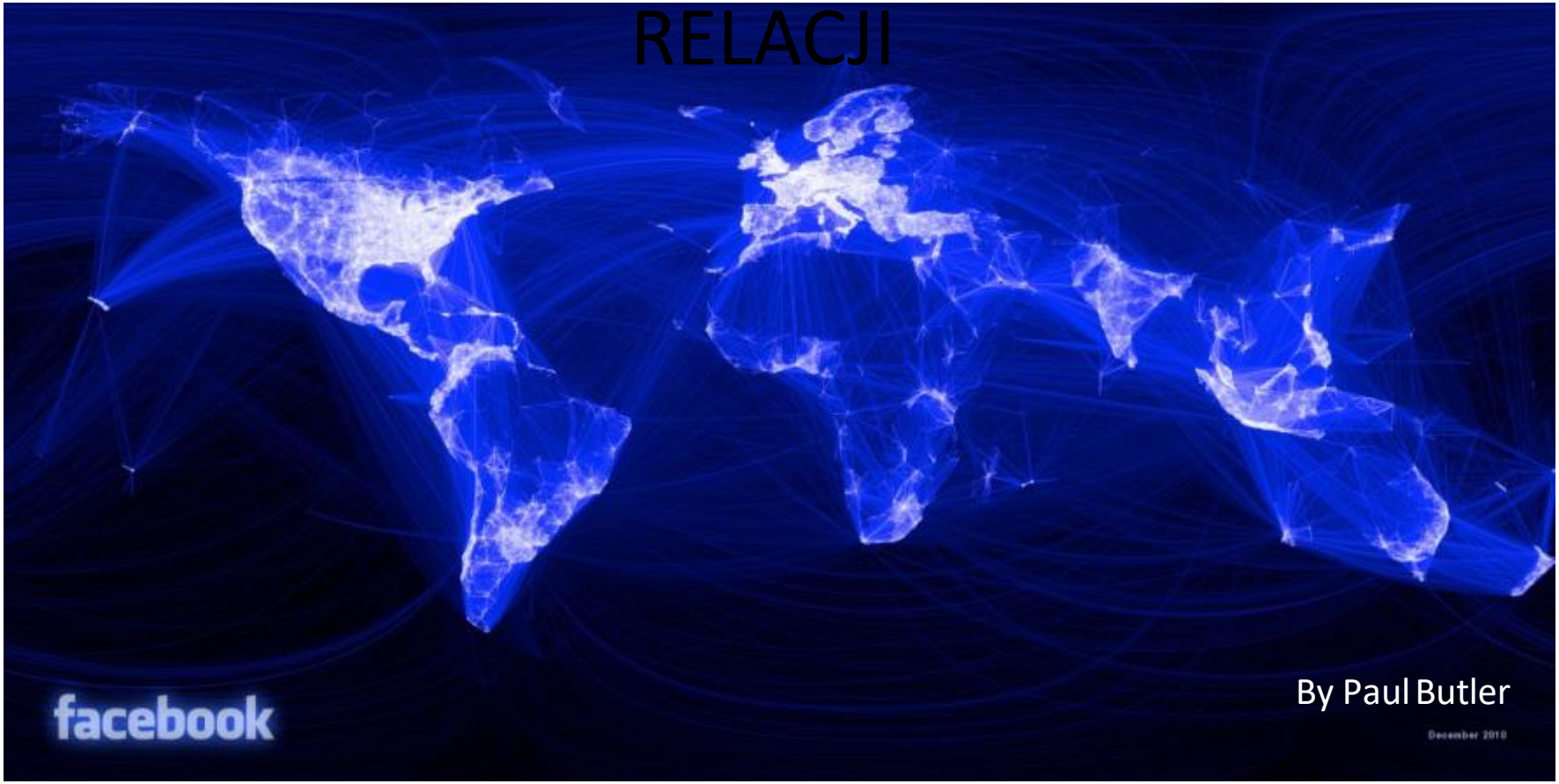
I wiele innych …

- **Sieci powiązań biznesowych**
- **Powiązania publikacji i cytowania**
- **Połączenia mózgowe**
- **Sieci komputerowe**
- **Sieć WWW**
- **Sieci powiązań biologicznych**

Powiązane tematy i kierunki badawcze

- Analizy statystyczne sieci
- Wizualizacja struktur sieciowych
- Teoretyczne modele sieciowe
- Sieci wielowarstwowe
- Sieci dynamiczne
- Rozprzestrzenianie informacji w sieciach
- Analizy rzeczywistych zbiorów danych

DUŻE SIECI SPOŁECZNE ➔ MILIARDY WĘZŁÓW I POWIĄZAŃ. RÓŻNE TYPY WĘZŁÓW I RELACJI



Celem wprowadzenia nie jest przedstawienie wszystkich aspektów tematu. Mogło by to być przedmiotem kompletnego cyklu wykładów. Głównie chodzi o wskazanie kluczowych metod, podejść, strategii i wybranych kierunków analiz i badań.

OBSZARY BADAŃ I ANALIZ

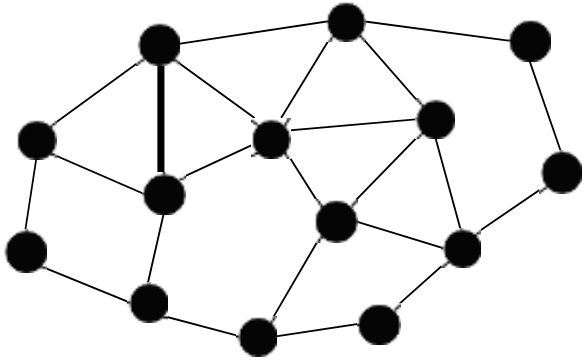
> 1.6 Miliarda

> 500 Milionów

> 200 Milionów

- Obserwacja całych sieci i zachodzących w nich zjawisk jest utrudniona lub wręcz niemożliwa.
- Załadowanie kompletnych złożonych struktur do pamięci komputera nie zawsze jest realne.
- Na tak dużych zbiorach danych problemem jest wyznaczenie nawet podstawowych miar sieciowych, długości ścieżek, średnicy sieci. Jeszcze trudniej wyznaczyć występujące wzorce czy struktury powiązań społecznych.

FACEBOOK -> NAJWIĘKSZA SIEĆ SPOŁECZNA



- 1 + miliard węzłów
- Średnio 130 linków wchodzących do każdego
- Węzła (średnio 130 znajomych)

> 1TB Pamięci, aby te relacje zapisać w postaci grafu, bez atrybutów, etykiet i treści

Pojawia się problem z ekstrakcją informacji, przetwarzaniem i analizami

Dwa możliwe rozwiązania: **Próbkowanie** i **Agregacja**

PRÓBKOWANIE VS AGREGACJA

- Próbkowanie sieci
  - Informacja o węzłach/linkach jest pozyskana dopiero po pobraniu próbki
  - Wymaga strategii eksploracji sieci i stopniowego powiększania próbki
  - Celem jest stopniowa identyfikacja małego zbioru **przedstawicieli** węzłów i powiązań ze struktury sieciowej, przy posiadanej niewielkiej wiedzy o całej sieci.
- Agregacja
  - Znana jest cała struktura sieci apriori
  - Celem są zagregowane miary, które umożliwią opis własności sieci na poziomie ogólnym, przy jak najmniejszej utracie informacji szczegółowych.

SIECI JEDNORODNE VS SIECI NIEJEDNORODNE

- **Homogeniczne → Single Relational Network**

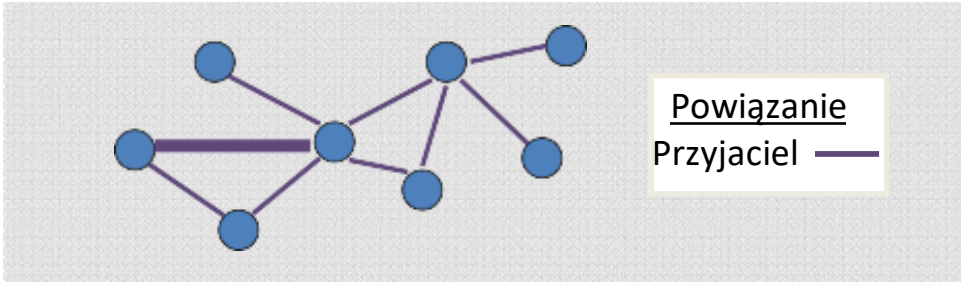
  Pojedynczy typ obiektu i typ linków

- **Heterogeniczne → Multi-Relational Network**

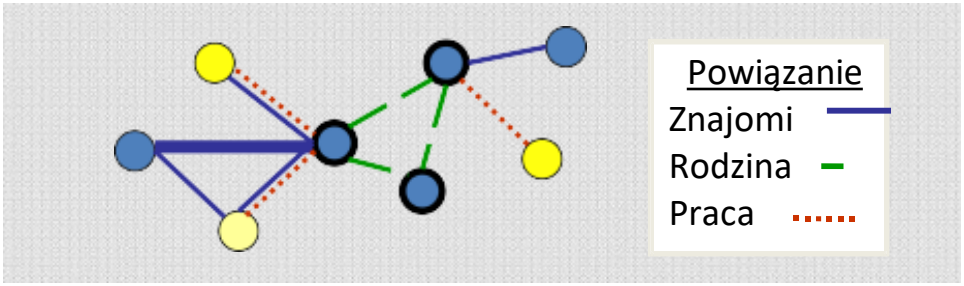
  Obiekty i linki **różnych typów**

Przykład

Jednorodne

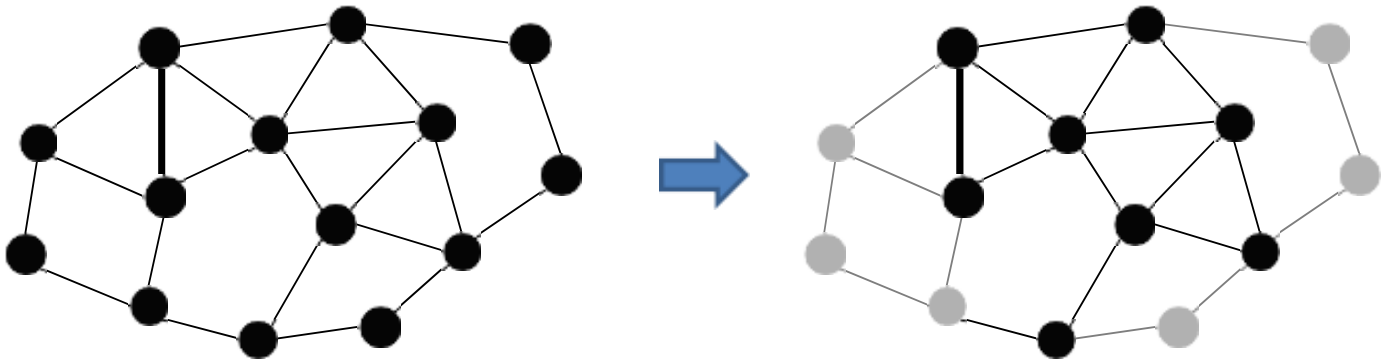


Niejednorodne



PRÓBKOWANIE SIECI

- Załóżmy, że szczegółowe informacje dotyczące węzła są dostępne dopiero w wyniku próbkowania. Struktura całej sieci nie jest znana.
- Celem jest mniejsza sieć, próbka, która powstaje w wyniku pobierania wycinkowych informacji. W zależności od zastosowanej metody zachowuje ona niektóre właściwości sieci pierwotnej.



OCENA JAKOŚCI SAMPLINGU

- W jaki sposób można mierzyć jakość próbkowania sieci?
- Próbkowanie możemy uznać za efektywne jeśli:
  - Uzyskana próbka zachowuje określone własności sieci
  - Analizy własności próbki sieci, na przykład analiza centralności, ścieżek, daje wyniki podobne do analiz własności kompletnej sieci
  - Uzyskana próbka jest znacznie mniejsza niż sieć pierwotna

ZACHOWANE WŁAŚCIWOŚCI PRÓBEK (1/3)

- Homogeniczne sieci statyczne
  - Rozkład stopni wierzchołka in/out
  - Rozkład długości ścieżek
  - Rozkład współczynnika klastrowania
  - Eigenvecor
  - Rozkład rozmiarów komponentów słabo i silnie połączonych
  - Struktura skupisk węzłów
  - I inne podobne

ZACHOWANE WŁAŚCIWOŚCI PRÓBEK (2/3)

- Jednorodne sieci dynamiczne
  - Dynamika gęstości sieci

    np. proporcja krawędzie vs węzły w czasie
  - Zmiany średnicy sieci w czasie

    np. zmniejszanie lub stabilizacja w czasie
  - Zmiany w czasie współczynnika grupowania
  - Wielkość macierzy sąsiedztwa
  - I inne podobne

ZACHOWANE WŁAŚCIWOŚCI PRÓBEK(3/3)

- Sieci heterogeniczne
  - Rozkład typów węzłów
  - Rozkład inter i intra linków łączących typy węzłów
  - Rozkład połączeń wyższego rzędu

METRYKI SIECIOWE

- Czy są zachowane właściwości sieci
  - Dla pojedynczych wartości np. współczynnik klastrowania, średnia długość ścieżki
  - Dla rozkładów właściwości np. rozkład stopnia, rozkład rozmiarów komponentów, odległości miedzy rozkładami miara np. KL divergence

- Realizacja procesów i zadań na samplach
  - Czy wyniki są podobne do zadań i procesów na sieciach kompletnych (np. procesy propagacji informacji, formowanie struktur i powiązań)

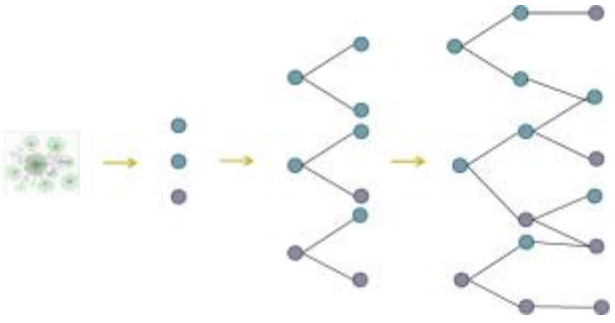
Próbkowanie sieci **homogenicznych**

DWIE GŁÓWENE STRATEGIE

- **Wybór węzłów lub krawędzi o zadanych ełaściwościach**

- **Pobierania próbek w procesie eksploracji**
  - Random Walk
  - Snow ball sampling
  - Poszukiwanie wzorców

Węzły początkowe (seeds)

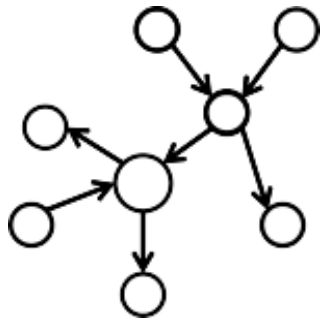


WYBÓR WĘZŁA

- Losowy wybór węzła
  - Losowy zbiór węzłów
- Wybór podstawie stopnia wierzchołka [Adamic, 2001]
  - Prawdopodobieństwo proporcjonalne do jego degeee wyboru węzła jest (zakładamy, że degree jest znane)
- PageRank sampling [Leskovec, 2006]
  - Prawdopodobieństwo wyboru węzła jest proporcjonalne do wartości jego miary PageRank (zakładając, że jest znana)

WYBÓR KRAWĘDZI

- Random Edge Sampling (RE)
  - Krawędzie wybieramy losowo, a następnie włączone są powiązane nimi węzły
- Random Node-Edge Sampling (RNE)
  - Wybieramy węzły a następnie powiązane z nimi krawędzie
- Hybrid sampling [Leskovec, 2006]
  - Z prawdopodobieństwem *p* realizowany jest RE sampling, a z prawdopodobieństwem *1 -p* RNE sampling



WYBÓR KRAWĘDZI

- Induced Edge Sampling [Ahmed, 2012]
  - Krok 1: Jednolity wybór krawędzi (a w konsekwencji węzłów) przez kilka rund
  - Krok 2: Dodawane są krawędzie, które są powiązane z wybranymi węzłami
- Frontier sampling [Ribeiro, 2010]
  - Krok 0: Losowo wybieraj zestaw węzłów L jako seeds
  - Krok 1: Wybierz element *u* z L przy użyciu degree based sampling
  - Krok 2: Wybierz krawędzie węzła u (u, v)
  - Krok 3: Zastąp u Przez v w zbiorze i dodaj (u, v) do sekwencji próbkowanych węzłów
  - Powtórz kroki 1 do 3

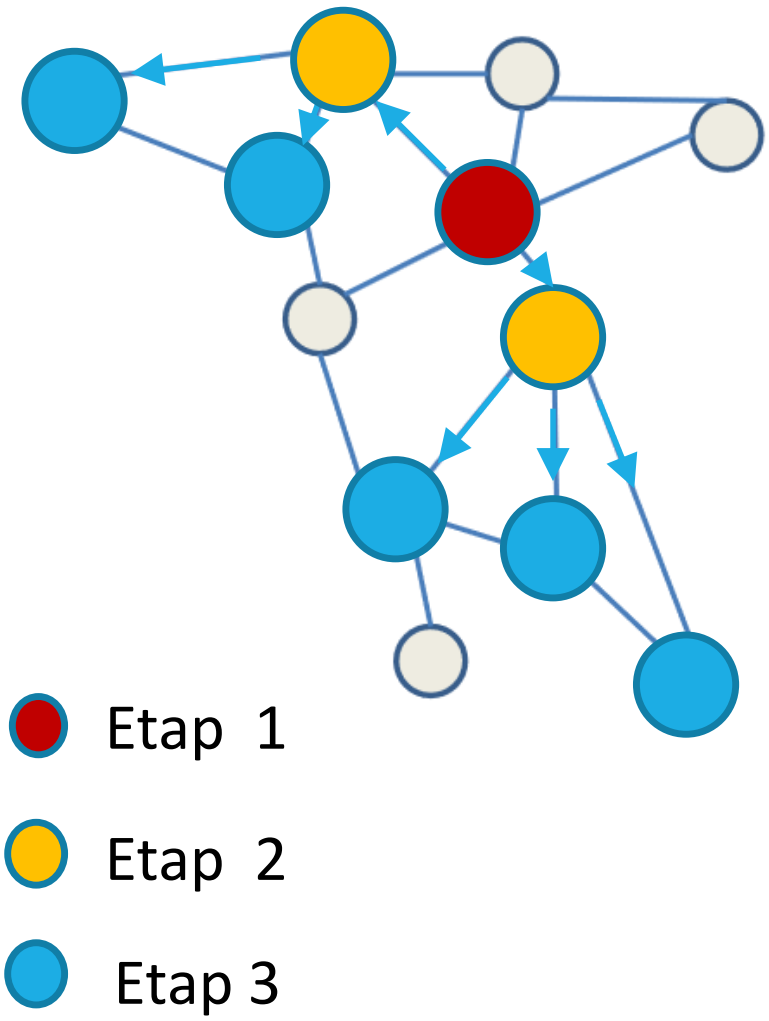
SAMPLING W PROCESIE EKSPLORACJI

- Snowball sampling

• Dla wskazanych początkowo węzłów, do próbki włączanych jest n sąsiadów wybieranych losowo. Proces postępuje iteracyjnie

• Węzły sąsiadujące są odwiedzane tylko wtedy, gdy nie zostały odwiedzone w poprzednich iteracjach

• Proces jest zrównoleglony, gdy w kroku pierwszym jest aktywowanych wiele węzłów

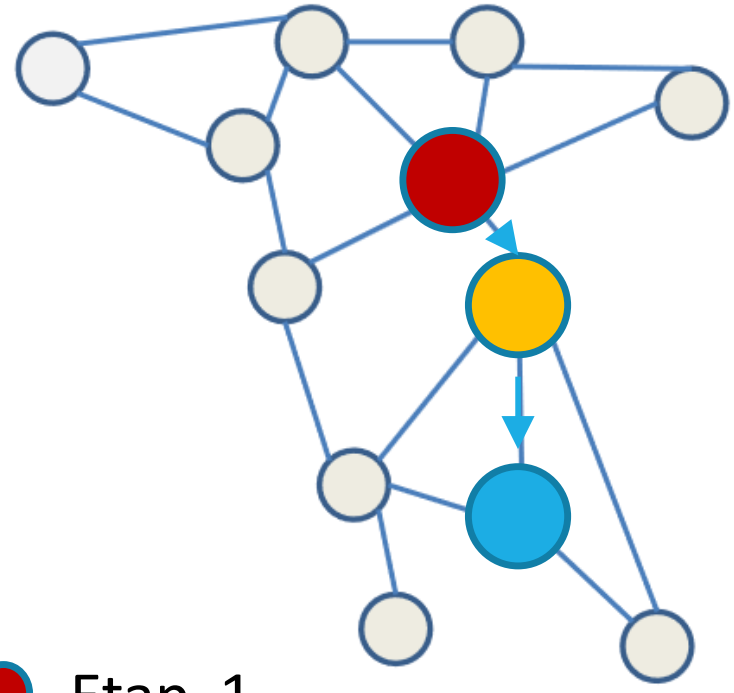


SAMPLING W PROCESIE EKSPLORACJI

- Random walk

- Dla aktywnego węzła jest wybierany losowo tylko jeden z jego sąsiadów i następuje do niego przejście.

- Próbka zawiera nadreprezentację węzłów z dużą liczba sąsiadów. Dla nich jest większe prawdopodobieństwo, że zostaną wybrane.



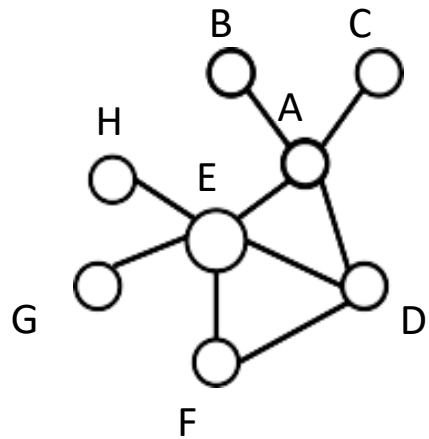
SAMPLING W PROCESIE EKSPLORACJI

- Modyfikacje Random walk [Gjoka, 2010]
  - Węzeł z następnego przeskoku jest wybierany jednolicie wśród sąsiadów bieżącego węzła
- Random walk z restartem [Leskovec ' 06]
  - Wybór węzła, random walk i ponowne uruchomienie
- Random jump [Ribeiro, 2010]
  - Podobnie jak random walk, ale z dodatkowo z prawdopodobieństwem p następują przeskoki do innych losowo wybranych węzłów sieci
- Forest fire [Leskovec, 2006]
  - Wybór węzła u
  - Losowe generowanie liczby z (<= liczba linków węzła u) i selekcja „z” linków jeszcze nie odwiedzonych
  - Krok wykonywany rekursywnie dla wszystkich nowo dodanych węzłów

SAMPLING W PROCESIE EKSPLORACJI

- Ego-centric exploration & sampling (ECE)
  - Zmodyfikowany random walk z przypisanymi prawdopodobieństwami selekcji uzależnionymi od właściwości węzła.
  - Multi ECE – rozpoczęcie procesu od **wielu seedów**
- Depth First/Breadth-First [Krishnamurthy, 2005]
  - Próbkowanie sąsiadów najczęściej odwiedzanych węzłów lub *ostatnio odwiedzanych*
- Sample Edge Count [Maiya, 2011]
  - Przekierowanie do sąsiada z najwyższym degree i kontynuacja od niego
- Expansion sampling [Maiya, 2011]
  - Konstruowanie próbek tak by maksymalizować ekspansję

SAMPLOWANIE EKSPANSYWNE

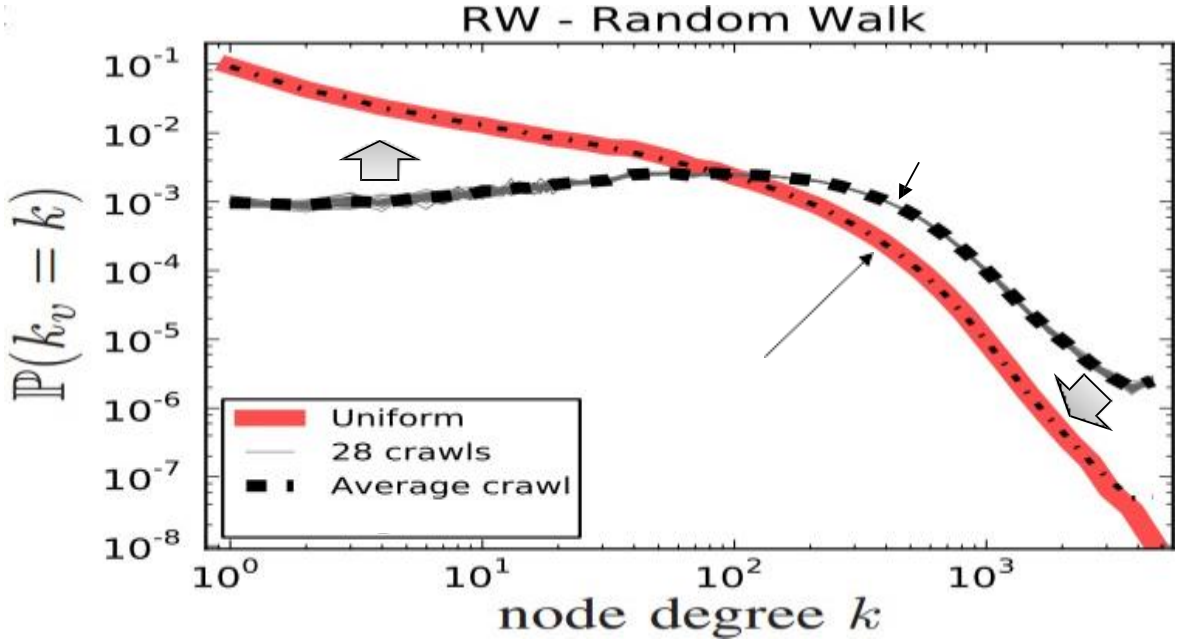


| N ({A}) |=4

| N ({E}) – N ({A}) ∪{A} | = | {F, G, H} | = 3

| N ({D}) – N ({A}) ∪{A} | = | {F} | = 1

RANDOM WALK = NADREPREZENTACJA HIGH DEGREE



q_K - próbkowany rozkład stopni

p_K - rozkład stopni pełnej sieci

- Średnie degree dla całej sieci ~ 94, średnie degree dla próbki ~ 338
- Rozwiązanie: modyfikowanie prawdopodobieństwa przejścia:

$$P_{v,w} = \begin{cases} \frac{1}{k_V} * \min\left(1, \frac{k_V}{k_w}\right) & \text{Jeśli w jest sąsiadem v} \\ 1 - \sum_{Y <> v} P_{v,y} & \text{Jeśli w = v} \\ 0 & \text{w przeciwnym razie} \end{cases}$$

METODA METROPOLIS

- Krok 1: Początkowo wybieramy próbkę S z losowo wybranymi n' węzłami
- Krok 2: Wykonywane iteracyjnie aż do konwergencji

  2.1 : Usuwamy jeden węzeł z S

  2.2 : Losowo dodajemy jeden węzeł do S → S'

  2.3 : Obliczamy współczynnik jakości

  $$a = \frac{\rho^*(S')}{\rho^*(S)}$$

  Jeśli a >= 1: akceptujemy $S := S'$

  Jeśli a < 1 : akceptujemy $S := S'$ z prawdopodobieństwem a

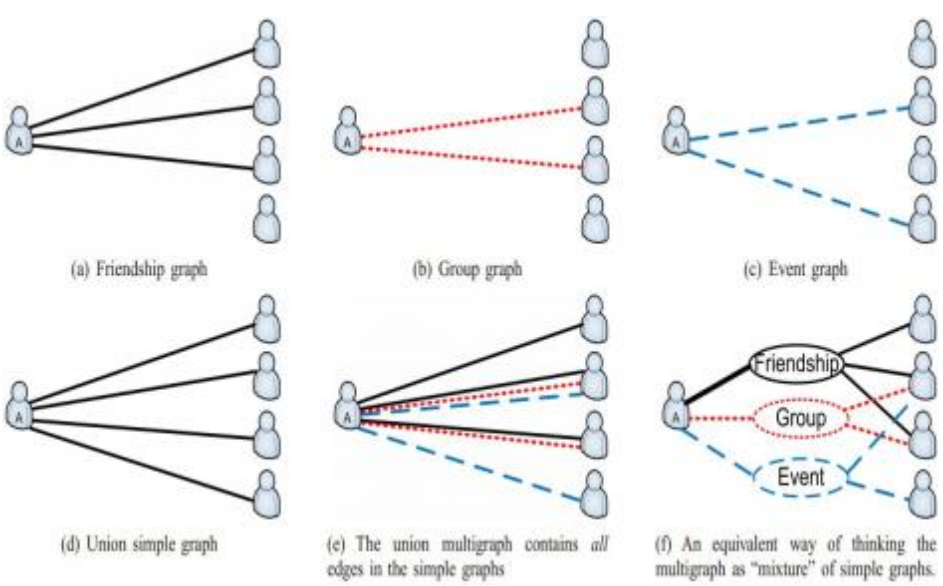
  odrzucamy $S := S'$ prawdopodobieństwem 1 - a

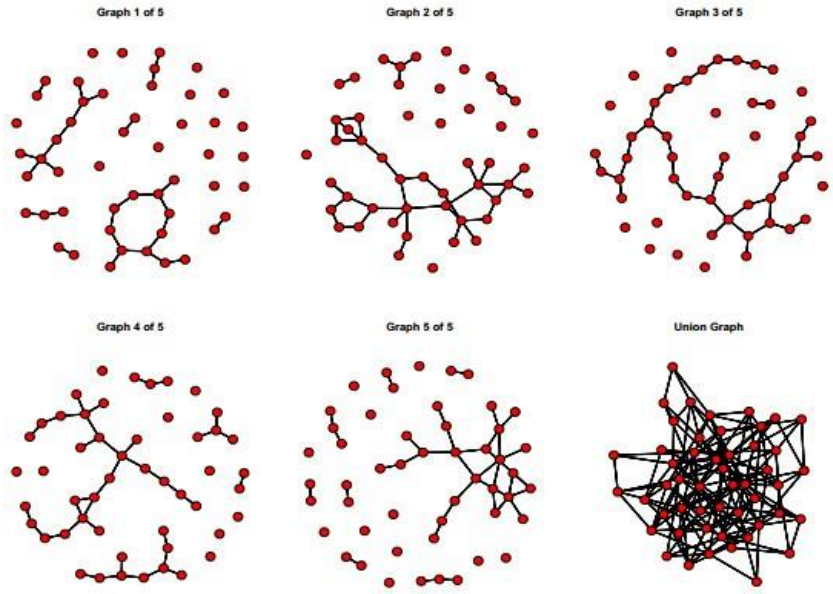
  – ρ^* (S) mierzy podobieństwo przyjętej własności między siecią S i siecią kompletną G
    - Może być uzyskane rozwiązanie przybliżone poprzez symulowane wyżarzanie

PRÓBKOWANIE SIECI HETEROGENICZNYCH

- Sieci niejednorodne
  - Graf G = <V, E> ma n węzłów ($v_1, v_2, ..., v_n$), m skierowanych krawędzi ($e_1, ..., e_m$) i k różnych typów
  - Każdy węzeł/krawędź jest przypisana do jednego z k typów typu $L = \{L_1, ..., L_K\}$
- Metody próbkowania HN
  - Multi-graph sampling [Gjoka,2010]
  - Pobieranie próbek z zachowaniem rozkładu typów [Li, 2011]
  - Próbkowanie z zachowaniem relacji [Yang, 2013]

MULTIGRAPH SAMPLING



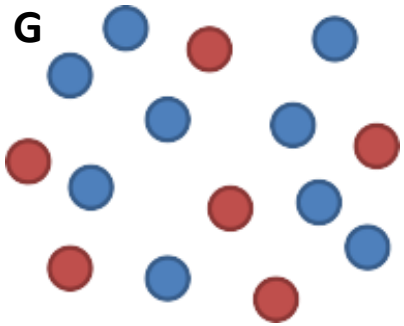


Random walk na multigrafie wynikowym, który powstaje jako rezultat unii grafów

PRÓBKOWANIE Z ZACHOWANIEM ROZKŁADU TYPÓW

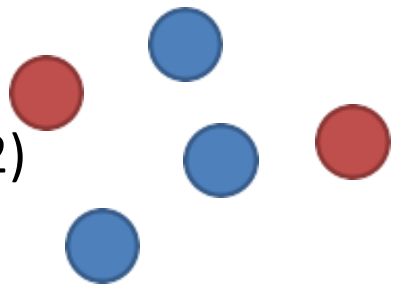
- Graf G i graf Gs utworzony w wyniku próbkowania
- Rozkład typów węzłów w grafie G_S opowinien być taki sam lub zbliżony do sieci pierwotnej G, $d(Dist(Gs),Dist(G)) = 0$
- $d()$ Oznacza różnicę pomiędzy dwoma rozkładami

Sieć pierwotna **G**



Próbka sieci **Gs**

(9:6) = (3:2)



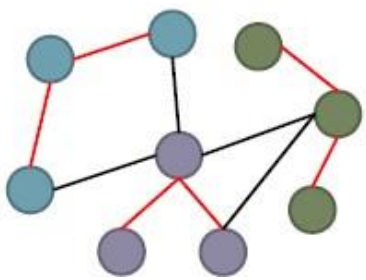
PRÓBKOWANIE Z ZACHOWANIEM TYPÓW POŁĄCZEŃ

- Połączenia heterogeniczne
  - Dla krawędzi $E[v_i, v_J]$
  - **Połączenie intra** - typ (v_I) = typ (v_J)
  - **Połaczenie inter** - typ (v_I) != typ (v_J)
- **Zachowanie relacji intra connection**

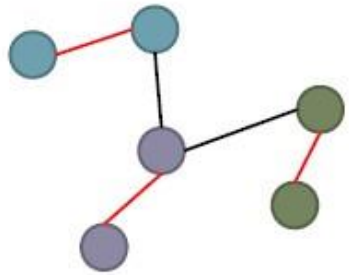
  Proporcje relacji wewnętrznych dla węzłów tych samych typów powinny być zachowane: $d\,(IR\,(G_S), IR\,(G)) = 0$
- Jeśli relacja wewnętrzna jest zachowana, to relacje zewnetrzna również jest zachowana

Próbka sieci **Gs**

Sieć pierwotna **G**

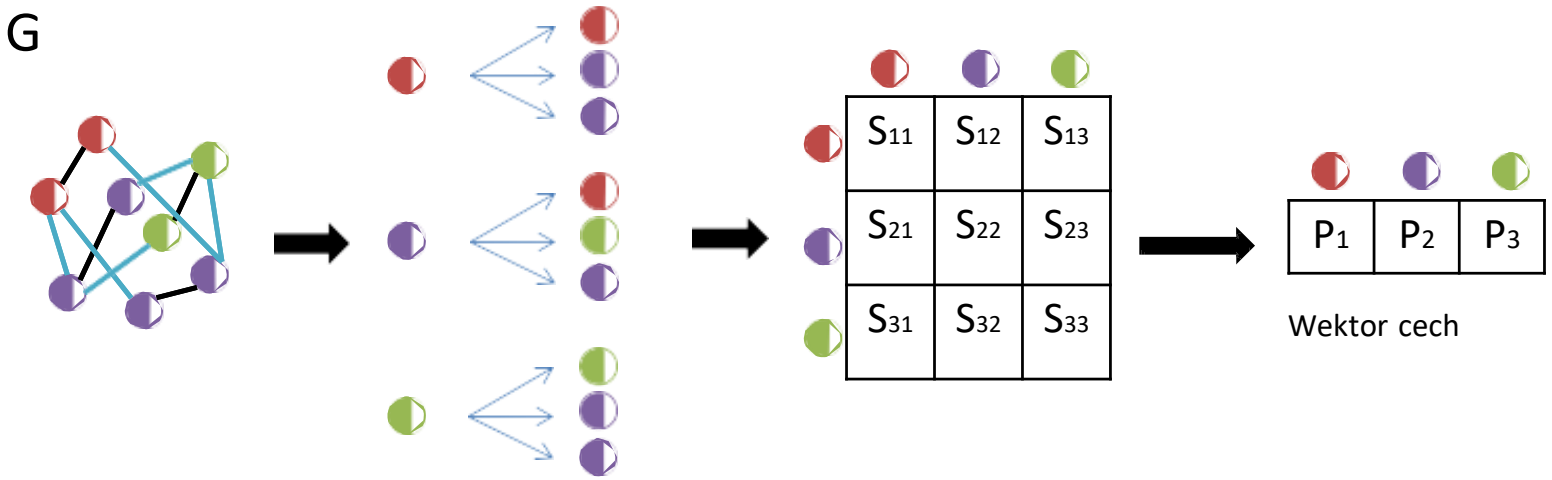


(6:4) = (3:2)

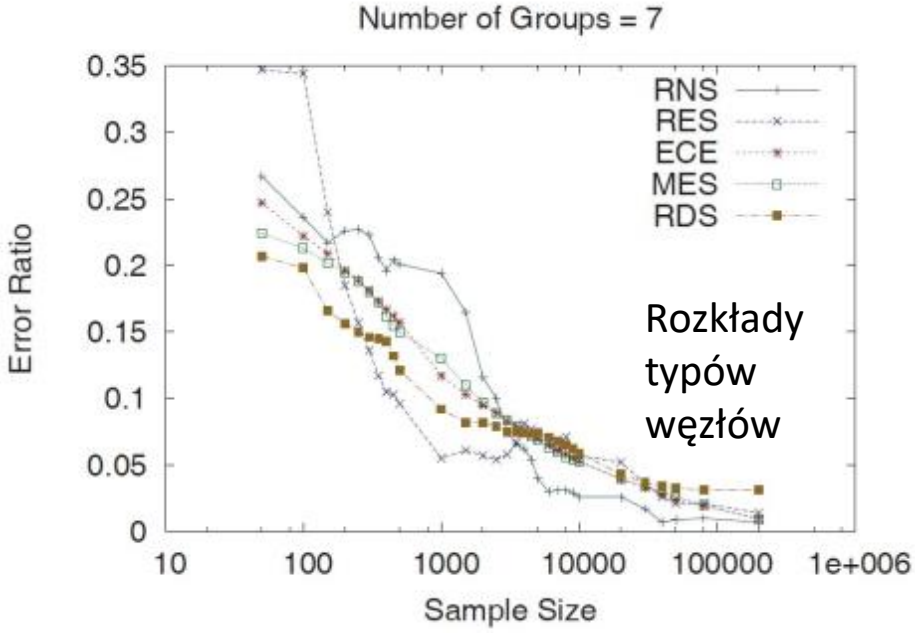


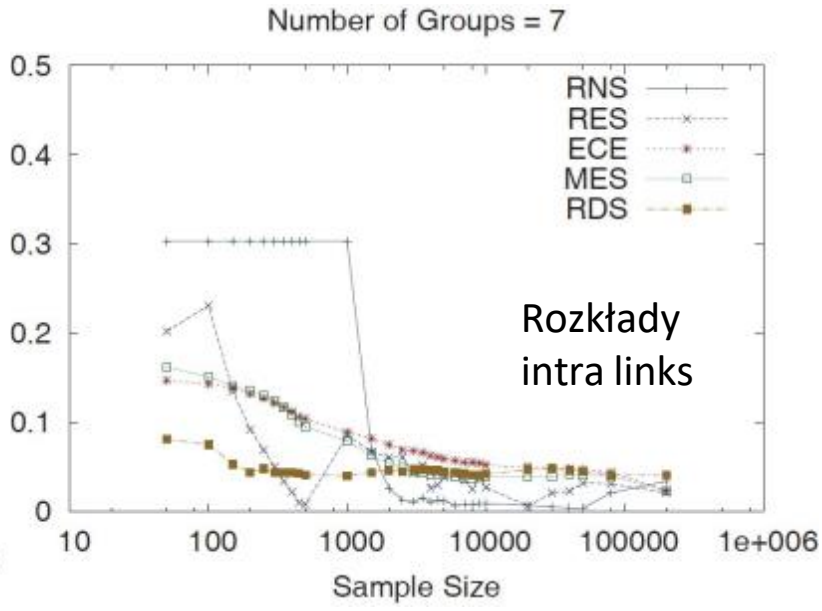
RESPONDENT-DRIVEN SAMPLING

- Zaproponowane dla badań ankietowych [Heck, 1999]
- Dwie główne fazy: próbkowanie Snowball ➔ poprawianie charakterystyk macierzy w celu lepszego dopasowania rozkładów



PORÓWNANIE RÓŻNYCH METOD PRÓBKOWANIA



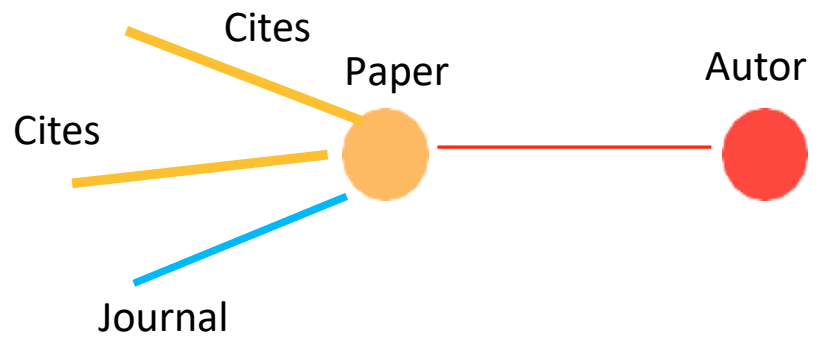


- Respondent driven sampling dobre efekty dla mniejszych próbek
- Próbkowanie losowe poprawia wyniki wraz ze wzrostem liczby węzłów

PRÓBKOWANIE Z ZACHOWANIEM RELACJI

- Zachowanie semantyki węzeł a nie struktury sieci
- **Profil relacyjny** uwzględnia równocześnie semantykę i strukturę powiązań
  - Brana jest pod uwagę zależność między typami węzłów i typami połączeń w sieci heterogenicznej
  - Składa się macierzy relacyjnych
    - Warunkowe prawdopodobieństwo $P(T_j | T_i)$ (np. P (ET = CITES | NT = paper))
    - węzeł - węzeł, węzeł - krawędź, krawędź - węzeł, krawędź - krawędź

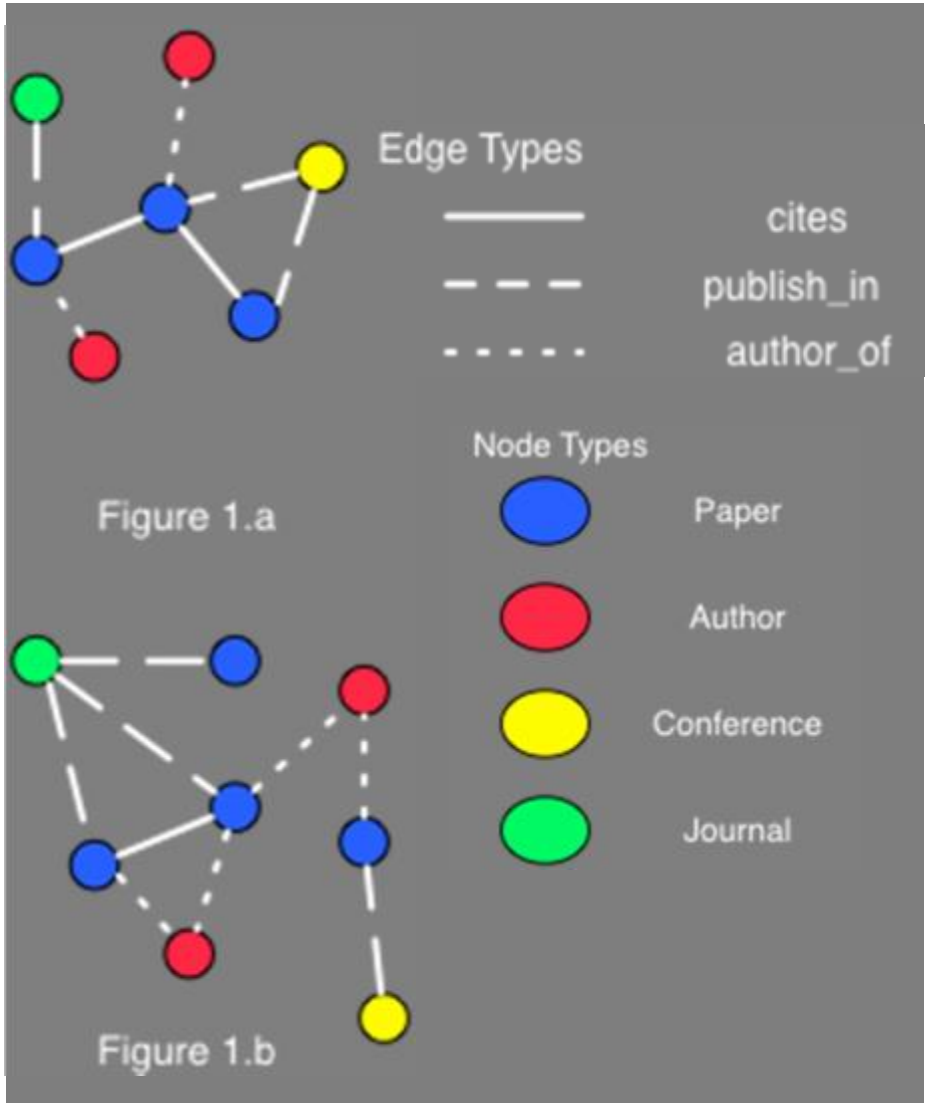
	Nt	Et
Nt	Macierz	Macierz
Et	Macierz	Macierz



PRZYKŁAD PROFILU RELACYJNEGO

	P	A	C	J	C	P	A
P	0.44	0.22 .	0,22	0,11	0.44	0.33	0,22
A	1						1
C	1					1	
J	1					1	
C	1				0,22	0,44	0,33
P	0.5		0,33	0,17	0,66		0,33
A	0.5	0,5			0.6.	0,4	

	P	A	C	J	C	P	A
P	0,182	0,364	0,091	0,273	0,182	0,364	0,364
A	1						1
C	1					1	
J	1					1	
C	1					0,5	0.5
P	0,5		0,125	0,375	0,17	0,5	0.33
A	0,5	0,5			0,22	0,33	0.44



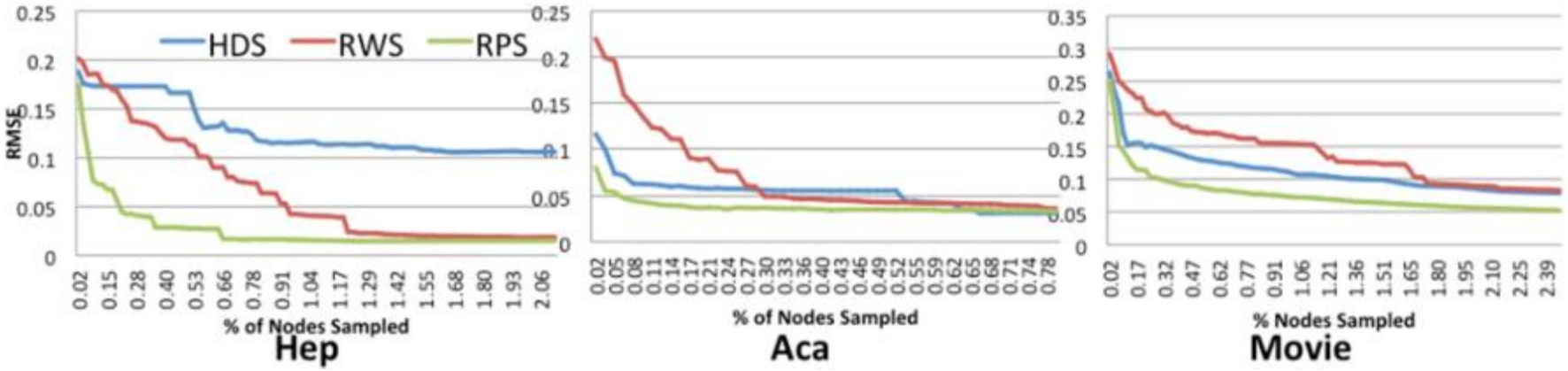
OCENA JAKOŚCI PRÓBKOWANIA

- Zbiory danych: trzy rzeczywiste sieci złożone
- Podstawowe metody jako punkty odniesienia:
  - Losowe próbkowanie Random Walk (RW)
  - Próbkowanie na podstawie degree (HDS)
- **Kryterium oceny I** (**Zachowanie właściwości**): w jaki sposób próbka aproksymuje własności całej sieci
- **Kryterium oceny II** (**Własności predykcyjne**): czy model predykcyjny uczony z wykorzystaniem próbki umożliwi predykcję nieznanych własności:
  - **Predykcja typu węzła**: przewidywanie typu węzła w sieci kompletnej na podstawie danych z próbki
  - **Przewidywanie brakujących relacji**: predykcja i odzyskiwanie brakujących linków
  - Funkcje:
    - in/out deg; avg in/out deg)
    - Jaccard's Coefficient
    - P(type(v)|Gs)= $\frac{\#type(v)=t \forall v \in N(n)}{|N(n)|}$
    - fRPnode = $\prod_{i \in N(n)} \frac{1}{Z} RP(type(i)|type(v)=t)P(type(v)=t)$
    - fRPpath = $\sum_{p \in Path(s,t)} \prod_{(p_1,p_2) \in p} P(type(p_2)|type(p_1))$

EKSPERYMENTY I ZACHOWANE WŁASNOŚCI

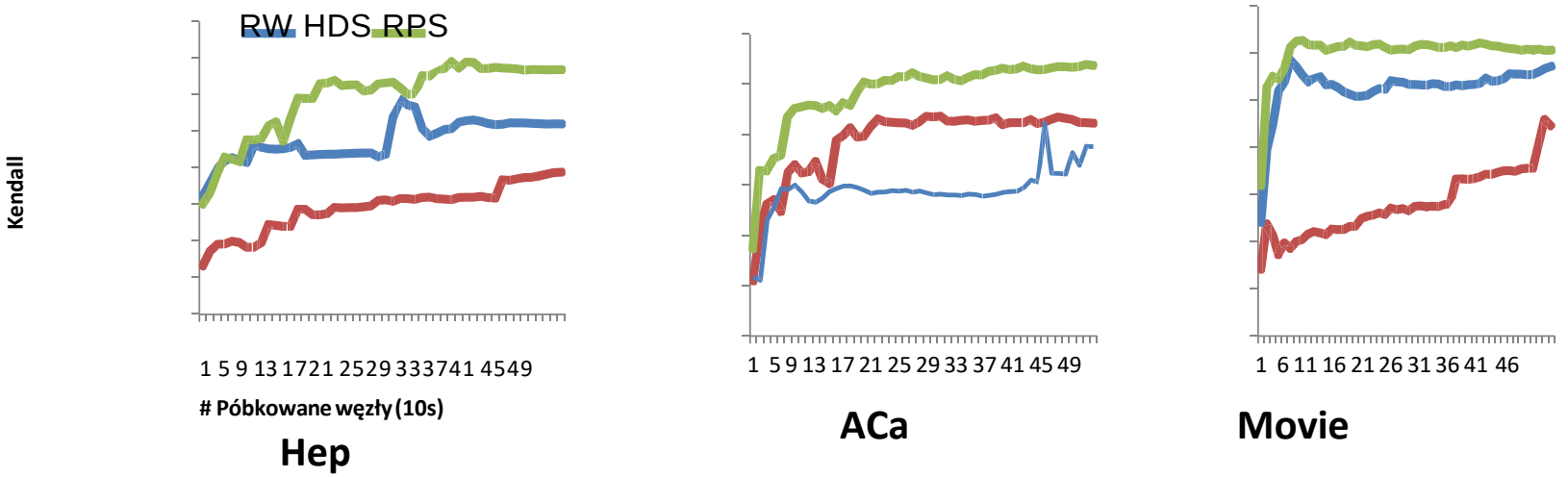
RP (RMSE)

Zachowanie typu zależności



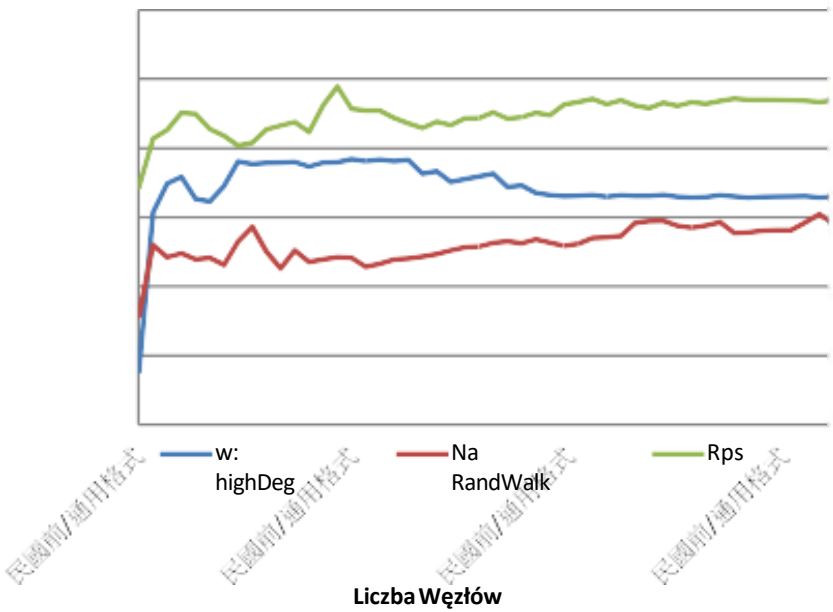
Ważony Pagerank

Zachowywanie wag węzłów

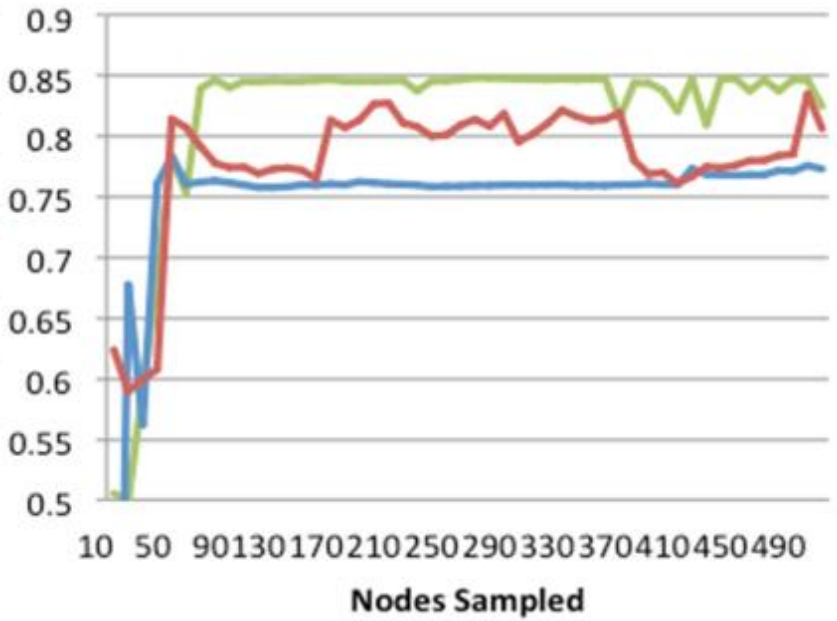


EKSPERYMENTY I PREDYKCJA

Przewidywanie typu węzła



Przewidywanie relacji



SAMPLING DEDYKOWANY

- Próbkowanie struktury społeczności [Maiya, 2010] [Satuluri,2011]
- Próbkowanie rdzenia sieci przy maksymalizacji wpływu [Mathioudakis, 2011]
- Pozyskiwanie danych na temat centralnych węzłów sieci [Maiya, ,2010]
- Predykcja rozkładu PageRank [Vattani, 2011]
- Predykcja linków [Ahmed, 2012]

PUBLIKACJE POWIĄZANE Z TEMATEM

	Sieci jednorodne	Sieci heterogeniczne
Wybór węzłów i krawędzi	[Leskovec,2006] [Adamic,2001] [Ahmed, rocznik, 2012] [Ribeiro, rocznik , 2010]	[Kurant, 2012]
Próbkowanie eksploracyjne	[Krishnamurthy, 2005] [Les wKoVEC, 2006] [Gjoka, 2010] [Ribeiro, 2010] [Maiya, 2011] [Kurant, 2011]	[Gjoka, 2011] [Li, 2011] [Kurant, 2012] [Yang, 2013]
Próbkowanie dedykowane	[Maiya,2010] [StulurI, 2011] [Mathioudakis, 2011] [Vattani, 2011] [Ahmed, 2012]	

Podsumowanie algorytmów i metod

Breadth/ Depth/ Random First Sampling (B-/D-/R- FS)
Snow-Ball Sampling (SBS)
Random Walk (RW)
Metropolis-Hastings Random Walk (MHRW)
Random Walk with Escaping (RWE)
Multiple Independent Random Walkers (MIRW)
Multi-Dimensional Random Walk (MDRW)
Forest Fire Sampling (FFS)
Respondent Driven Sampling (RDS) (RWRW)

Materiały źródłowe

Granovetter, M. (1976). Network sampling: Some first steps. *American journal of sociology*, *81*(6), 1287-1303.

Hu, P., & Lau, W. C. (2013). A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*.

Lin, S. D., Yeh, M. Y., & Li, C. T. (2013). Sampling and summarization for social networks. In *17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)(tutorial)*.

Biblioteki programistyczne

https://rdrr.io/cran/igraph/man/random_walk.html

http://www.michelecoscia.com/?page_id=1390

https://www.rdocumentation.org/packages/netdep/versions/0.1.0/topics/snowball.sampling