Coś z niczego?
czyli czego możemy się dowiedzieć o sieciach społecznych na podstawie danych sondażowych

Michał Bojanowski

Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
Uniwersytet Warszawski

Jabłonna, 28 września 2016

Problem

Badania sieci społecznych opierają się zazwyczaj na

1. badaniu wyczerpującym (census sieciowy, trudne, drogie)
2. badanie sieci ego-centrycznych (losowa próba węzłów, mniej trudne, tańsze)

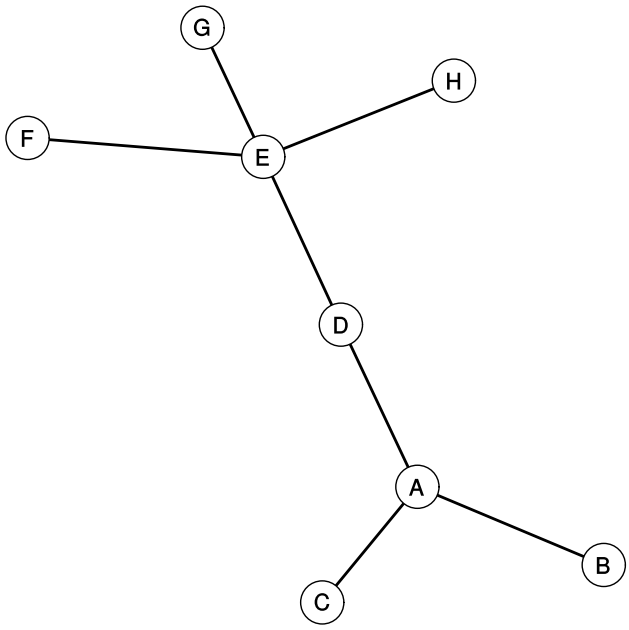
Czy, kiedy i z jakim skutkiem jesteśmy w stanie (1) zastąpić (2)?

W dalszej części

- Zbieranie danych: census a sieci-egocentryczne
- ERGM, w **wielkim** skrócie, dla kompletnych danych sieciowych
- ERGM dla danych ego-centrycznych – symulacja
- ERGM dla danych ego-centrycznych – ilustracja na podstawie danych ogólnopolskich (robocza)

Sieci społeczne?

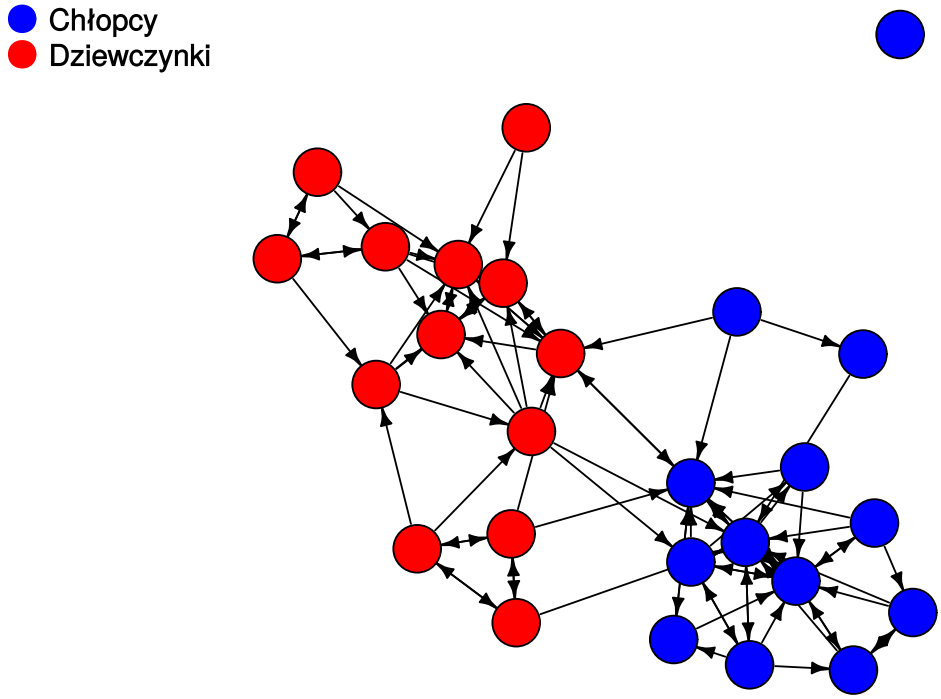
- węzły
- relacje
- atrybuty węzłów
- atrybuty relacji



Zbieranie danych sieciowych

- **Badanie wyczerpujące**: posiadamy informacje o wszystkich węzłach i łączących ich relacjach.
- **Schemat egocentryczny**: bazujemy na (losowej) próbie aktorów i raportowanych przez nich informacjach o ich bezpośrednim otoczeniu sieciowym.
- **Schemat podążania za relacjami** (*link-tracing*): Szeroki zbiór metod typu "kuli śniegowej".

Z kim chciałbyś się bawić?



Źródło: IBE, SUEK

?

- Dlaczego sieć wygląda tak, jak wygląda?
  - *Odwzajemnianie*
  - *Zróżnicowanie dzieci ze względu na popularność (relacje przychodzące)*
  - *Zróżnicowanie dzieci ze względu na "towarzyskość" (relacje wychodzące)*
  - *Homofilia / segregacja: dzieci chcą się bawić z innymi tej samej płci*
  - *Przechodniość*

Exponential-family Random Graph Models

- Model probabilistyczny
- Rozkład prawdopodobieństwa na zbiorze wszystkich grafów o zadanej liczbie wierzchołków

$$P(Y = y) = \frac{\exp \theta g(y)}{k(\theta)}$$

Gdzie:

- $P(Y = y)$ prawdopodobieństwo zaobserowowania grafu y
- $g(y)$ wektor statystyk sieciowych, np. liczba relacji, liczba relacji odwzajemnionych, przechodnie tryplety itd.
- θ wektor parametrów
- $k(\theta)$ stała normalizująca aby prawdopodobieństwa sumowały się do 1.
- Statystyki $g(y)$ związane z częstością występowania różnego rodzaju (małych) konfiguracji relacji i/lub atrybutów węzłów i relacji

Postać logit

$$\log\left(\frac{P(Y_{ij}=1|y_{ij}^c)}{P(Y_{ij}=0|y_{ij}^c)}\right) = \text{logit}(Y_{ij}=1|y_{ij}^c) = \theta'\delta(y_{ij})$$

Gdzie:

- $P(Y_{ij}=1|y_{ij}^c)$ warunkowe prawdopodobieństwo, że i i j pozostają w relacji przy warunku, że reszta grafu (dopełnienie) pozostaje niezmieniona.
- y_{ij}^c dopełnienie diady ij – wszystkie diady oprócz y_{ij}
- $\delta(y_{ij})$ zmiana wartości statystyki sieciowej g na skutek zmiany stanu relacji y_{ij}

$$\delta(y_{ij}) = g(y_{ij}^+) - g(y_{ij}^-)$$

Przykłady statystyk sieciowych ($g(y)$)

- Gęstość (liczba relacji) $\sum_{ij} y_{ij}$
- Efekty związany z atrybutami węzłów (grupowe różnice w średniej liczbie relacji)
- Efekt związany z atrybutami diad (np. homofilia)
- Stopień (*degree*)
- Liczba trójkątów $\sum_{i<j<h} y_{ij} y_{ih} y_{jh}$
- …

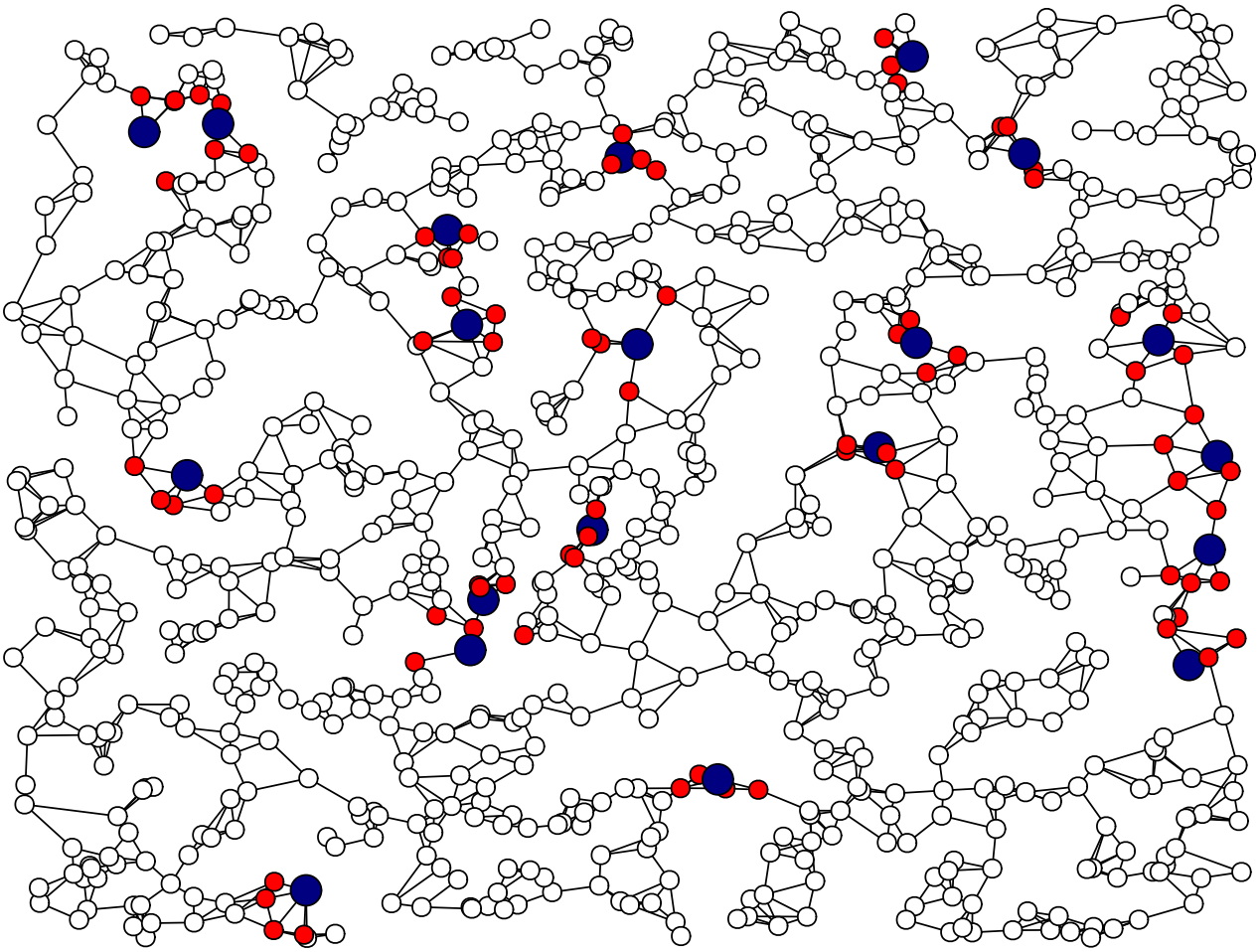
Przykładowy model

	Estimate	Std. Error	p-value
Liczba relacji	-2.963	1.507	0.050
Dziewczynki (towarzyskość)	0.750	0.402	0.063
Dziewczynki (popularność)	0.167	0.411	0.686
Płeć (homofilia)	2.481	0.445	0.000
Status (towarzyskość)	0.197	0.303	0.517
Status (popularność)	0.016	0.313	0.959
Status (homofilia)	0.286	0.308	0.354
IQ (towarzyskość)	-0.002	0.012	0.866
IQ (popularność)	-0.012	0.011	0.289
Przechodniość	0.916	0.464	0.049

Widzimy m.in.:

- Dziewczynki są bardziej towarzyskie od chłopców
- Homofilia ze względu na płeć: relacje pomiędzy dziećmi tej samej płci są $e^{2.481} = 11.953$ razy bardziej prawdopodobne niż relacje pomiędzy dziećmi różnych płci.
- "Znajomy znajomego jest moim znajomym"
- Zbyt mało danych by powiedzieć coś konkretnego o roli pozostałych czynników.

Schemat egocentryczny



ERGM na podstawie danych egocentrycznych

Założenia:

- Losowa próba ego-sieci
- Alters nie są jednoznacznie identyfikowalni: abtrahujemy od sytuacji w której alter może też być ego, albo ta sama osoba jest alterem dwóch ego.
- Próba ego-sieci jest infinityzymalną frakcją populacji

Powyższe założenia ograniczają rodzaj statystyk sieciowych, które możemy użyć w ERGM

- Liczba relacji, efekty atrybutów węzłów i relacji, stopień = tak
- Przechodniość i inne triadowe = nie
    - *(chyba, że pytamy ego o relacje alter-alter)*

Ego-statystyki wymagają przeskalowania

- np. liczba relacji w ego sieciach jest dwukrotnie wyższa niż w badaniu wyczerpującym, bo każda relacja jest raportowana dwa razy (przez oba połączone węzły)
- itp.

Symulacja

Na ile dobrze można oszacować parametry na podstawie danych ego-centrycznych?

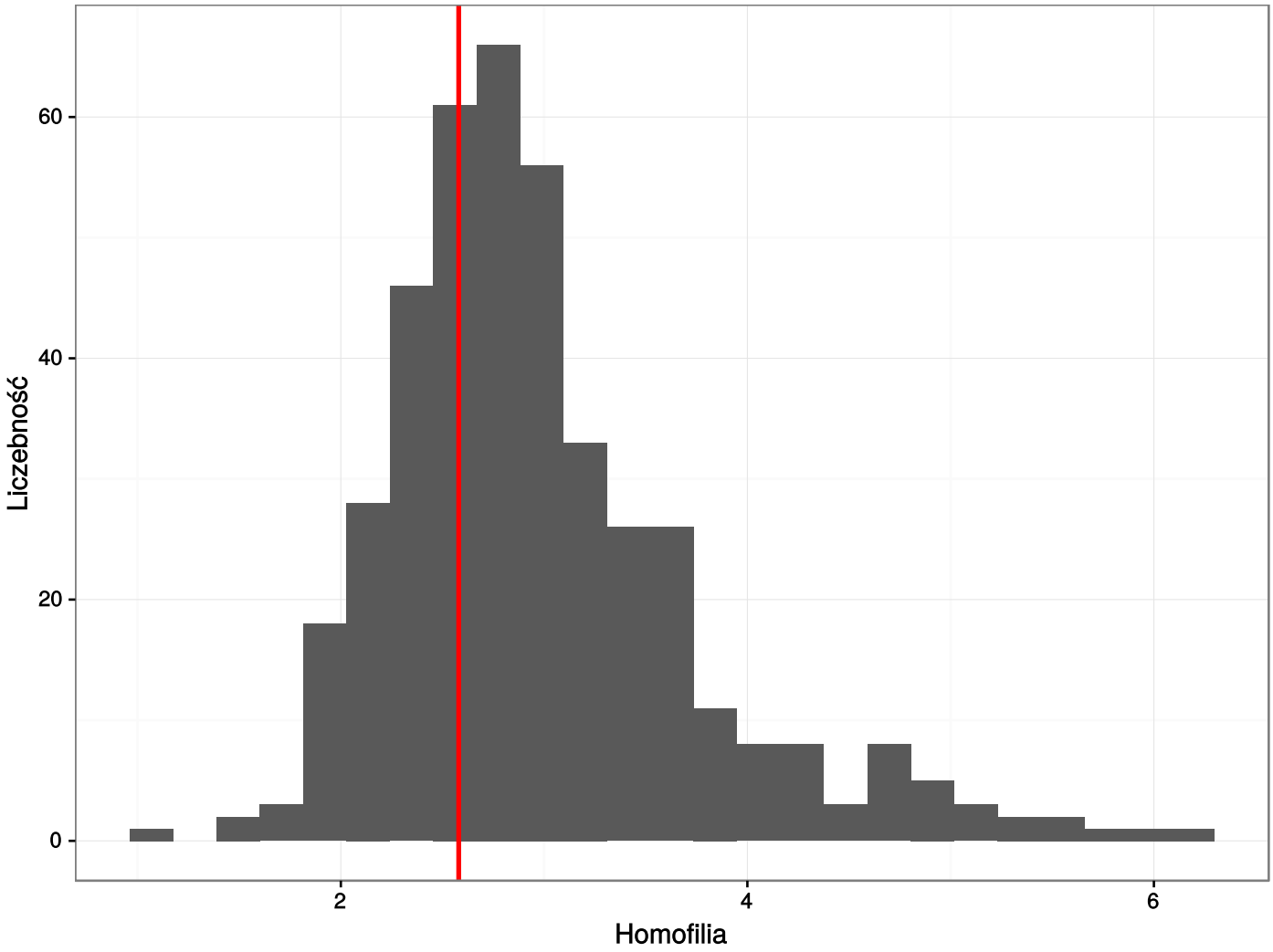
Konstrukcja:

- Badamy dzieci w klasie (dane IBE powyżej)
- Wielkości prób jako frakcja populacji: 0.3, 0.45, 0.6, 0.75, 0.9
- Dla każdej wielkości próby 100 różnych prób, razem 500 sieci (5×100)
- W każdej próbie szacujemy model z efektami:
  - *liczba relacji*
  - *homofilia ze względu na płeć*

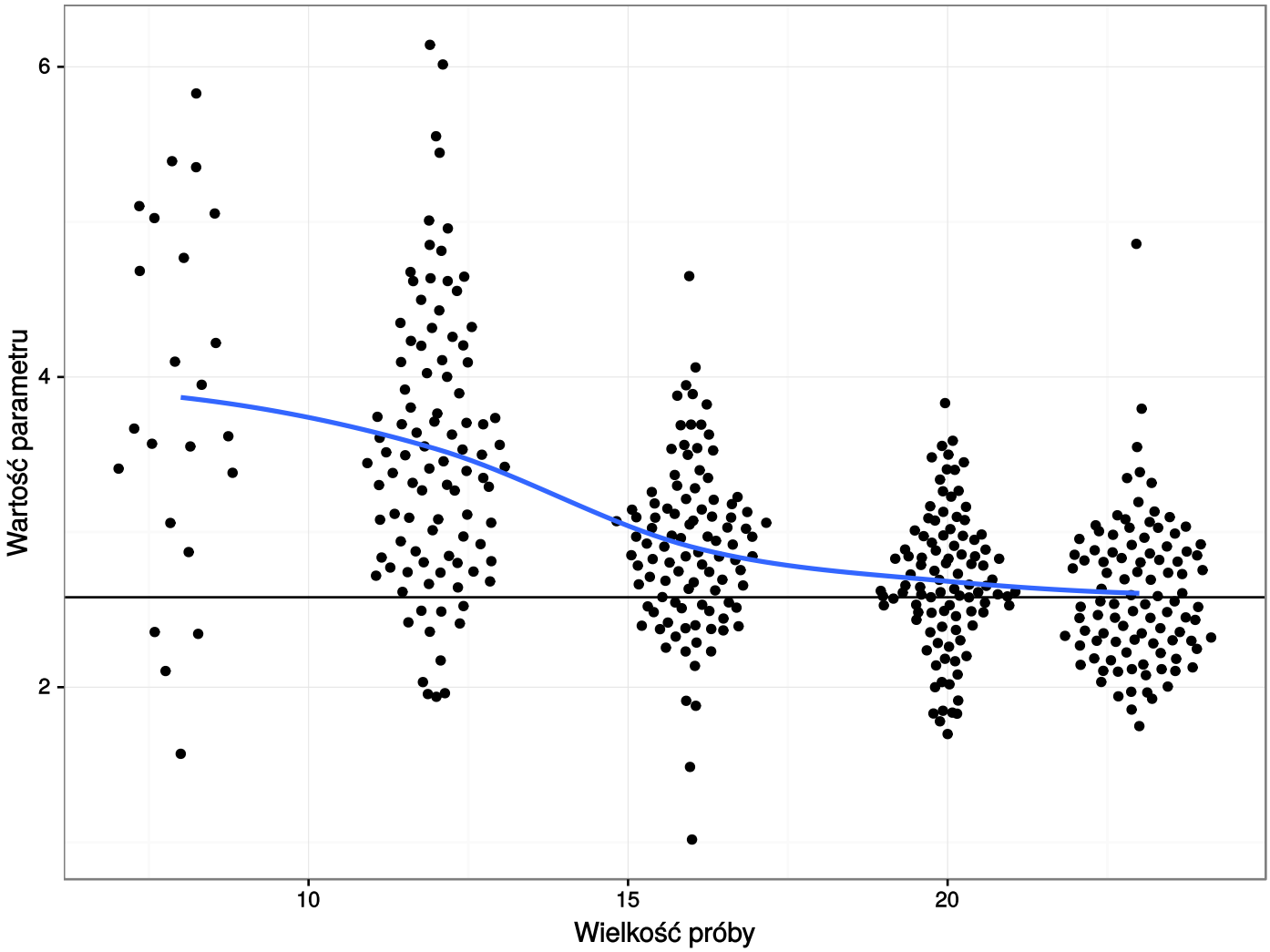
Pytania:

- Na ile szacunki będą zgodne z wynikami modelu na danych pełnych?
- Jak szacunki zależą od wielkości próby?

Zróżnicowanie wyników pomiędzy próbami



… w zależności od wielkości próby



Obserwacje

- Wyniki z ego-sieci są obciążone
  - *Próbkujemy węzły by dowiedzieć się o relacjach*
  - *Duża homofilia ze względu na płeć*
- Oszacowania są zgodne (*consistent*) - im większa próba, tym bliżej prawdziwej wartości
- Znaczna wariancja z próby

Ogólnopolski sondaż sieci personalnych

Dane

Projekt *Ludzie w sieciach: Wpływ kontekstu społecznego na jednostkę i jego rola w kształtowaniu struktury społeczeństwa* (kier. Bogdan W. Mach, NCN HS6/00526).

- Ogólnopolska próba PESEL
- "Generator imion"
- Wykorzystane zmienne o ego i alter: wiek, płec, wykształcenie

Model

ERGM z efektami głównymi oraz homofilii dla płci, wieku oraz wykształcenia.

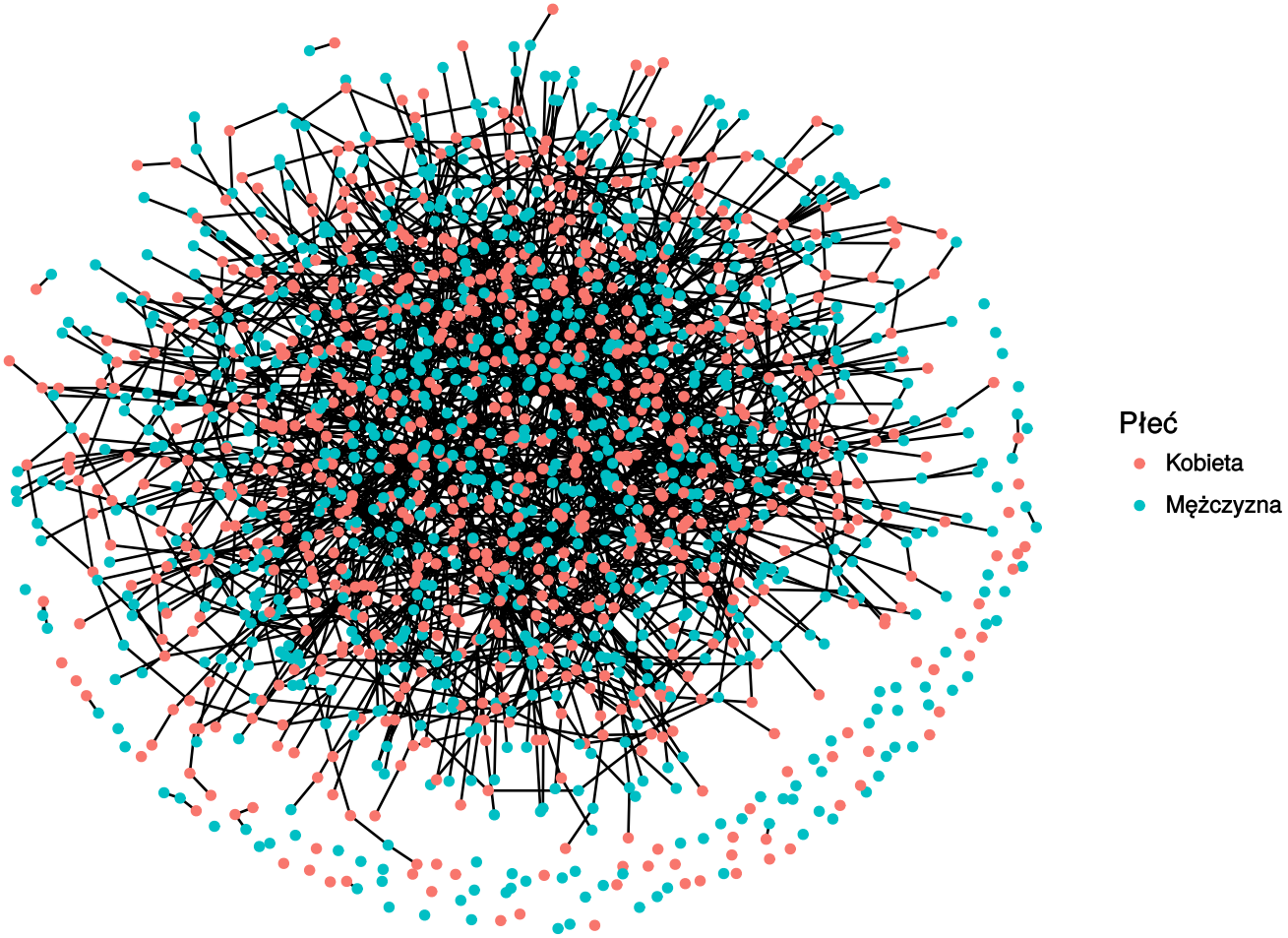
Model (wyniki)

	Estimate	Std. Error	p-value
Offset	-7.367	0.000	0.000
Liczba relacji	3.561	0.538	0.000
Płeć (homofilia)	-1.455	0.042	0.000
Płeć (mężczyźni)	-0.242	0.044	0.000
Wykształcenie (homofilia)	1.105	0.048	0.000
Wykształcenie (zawodowe)	-0.201	0.093	0.030
Wykształcenie (średnie)	-0.170	0.096	0.077
Wykształcenie (wyższe)	-0.039	0.109	0.720
Wiek (18-29)	-1.341	0.242	0.000
Wiek (30-39)	-1.326	0.234	0.000
Wiek (40-49)	-1.347	0.243	0.000
Wiek (50-59)	-1.468	0.238	0.000
Wiek (60-69)	-1.421	0.250	0.000
Wiek (70<)	-1.008	0.261	0.000
Wiek (homofilia)	1.767	0.049	0.000

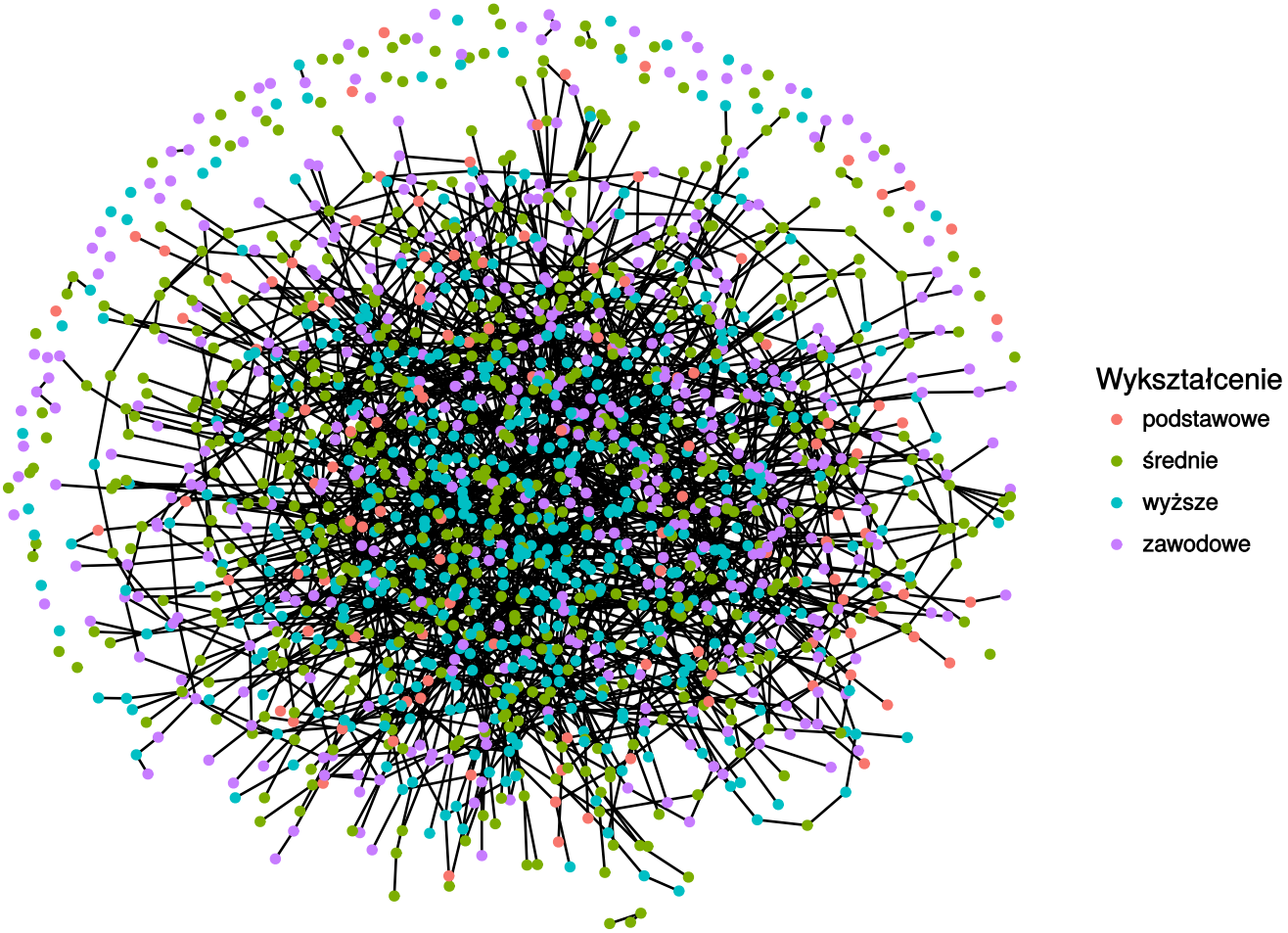
Model (spostrzeżenia)

- Mężczyźni średnio wymieniają mniej osób
- **Heterofilia** ze względu na płeć (!?)
- Homofilia ze względu na wykształcenie
- Liczba znajomych wydaje się maleć z wiekiem
- Homofilia ze względu na wiek

Sieć (płeć)



Sieć (wykształcenie)



Podsumowanie

- Dane ego-centryczne zawierają bogate informacje na temat struktury sieci jako całości
- Informacja ta oraz ERGM-y mogą posłużyć do
  - *Oszacowania modelu dla sieci jako całości*
  - *Symulacyjne generowanie sieci posiadających własności zgodne z obserwowanymi w ego-sieciach*
  - *Badanie procesów zachodzących w sieciach na danych symulacyjnych*
- Problemy:
  - *Specyfikacja modeli ograniczona do efektów/statystyk "dostępnych" w danych ego-centrycznych*
  - *Estymatory wydają się być obciążone (przynajmniej w przypadku homofilii)*
  - *Estymatory mają znaczącą wariancję*

Literatura

- Krivitsky, P. N., Handcock, M. S., & Morris, M. (2011). Adjusting for network size and composition effects in Exponential-family Random Graph Models. *Statistical methodology*, 8(4), 319-339.
- Krivitsky, P. N., & Morris, M. (2015). *Inference for Social Network Models from Egocentrically-Sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the US*. Working paper 05-15, University of Wollongong.

Dziękuję!

Michał Bojanowski

m.bojanowski@uw.edu.pl