

Hybrid Probablistic-Snowball Sampling Design

**I.e. Weird network crawling w/ attrition,
by Giulio G. Cantone**

Why this work is supposed to be relevant?

Year cost per continent to monitor smoking prevalence: ~7M €

Smoking-prevalence: $\frac{n(smokers)}{N(pop)}$ Lately, a bit fuzzy: is a vaper a smoker?

Why this cost: you want a good randomisation strategy.

Randomisation = Probabilistic Sampling a.k.a **Gold Standard** Design

(E.g. you use **Gold Standard** for calculating Design Effects for advanced estimator)

When you are a poor social scientist, you don't randomise

Students of psych/sociol/mkting just code a questionnaire and ask to their friends to share it through their social media.

We can call this “Fake Snowball Sampling Design”.

This is widely recognized as **not scientific at all**
and results of it are of little scientific interest and value...

...however...

...however...

Historically there are at least 2 sampling designs that use social connections with have an analytical model behind:

- True Snowball Sampling ('60)
- Respondent-Driven Sampling ('90)

Both have been employed to sample attribute-prevalence within **hiding populations**: political extremists, drug abusers, HIVs...

RESEARCH QUESTION

Are there sufficient conditions to employ the knowledge developed from Snowball/RDS for decreasing the costs of Gold Standards surveys in **non-hiding** populations?

E.g. smoking prevalence?

HYBRID PROBABILISTIC-SNOWBALL SAMPLING DESIGN (HPSSD) will try to answer this question

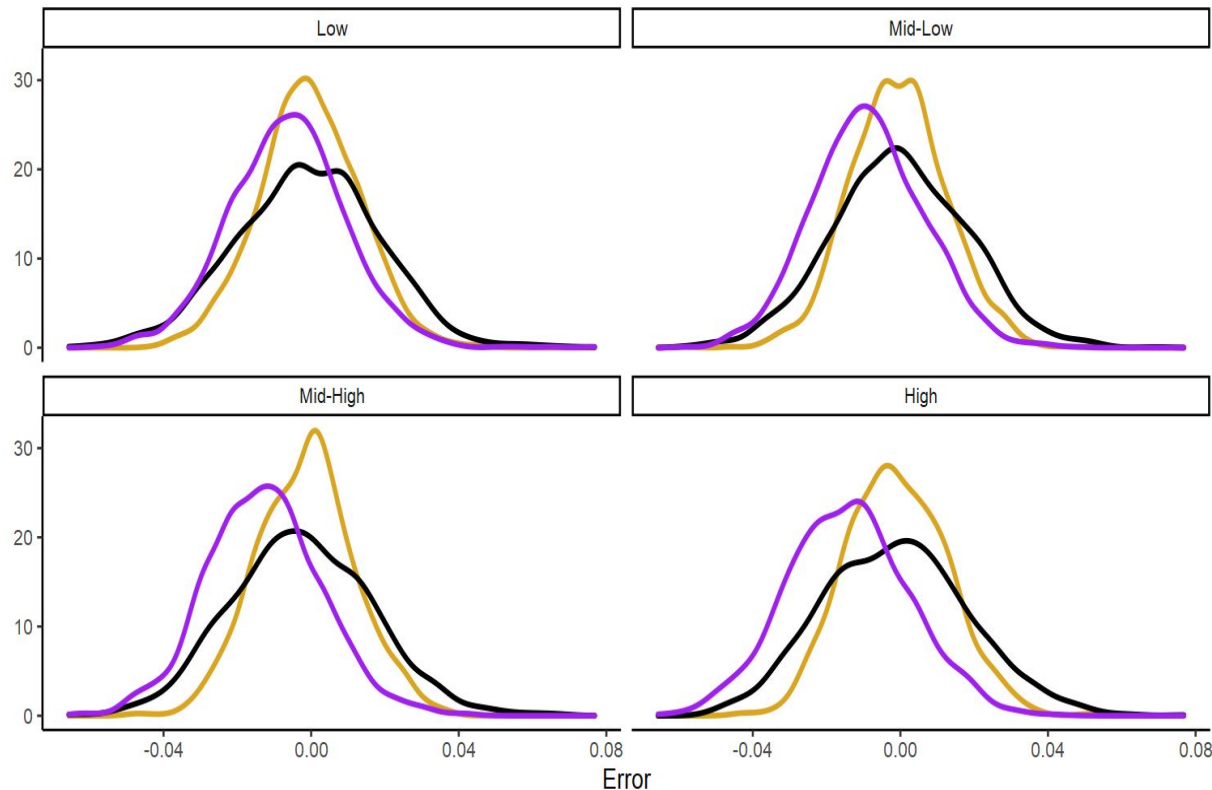
VISUALIZING THE QUESTION WE WANT TO ANSWER

We want: errors like the **Golden Curve**.

Would you pick:
Black? No Bias, more Variance
Purple? Bias, less Variance

Low, High... etc are levels of degree-homophily.

Results are work-in-progress.



A TRICK TO REMEMBER KURTOSIS VS BIAS: CHANGE OF PERSPECTIVE

LEPTOKURTOSIS OF ERRORS -> HIGH PRECISION

ABSENCE OF BIAS IN ERRORS -> HIGH ACCURACY

Accurate
Precise



Not Accurate
Precise



Accurate
Not Precise



Not Accurate
Not Precise



GOOD!

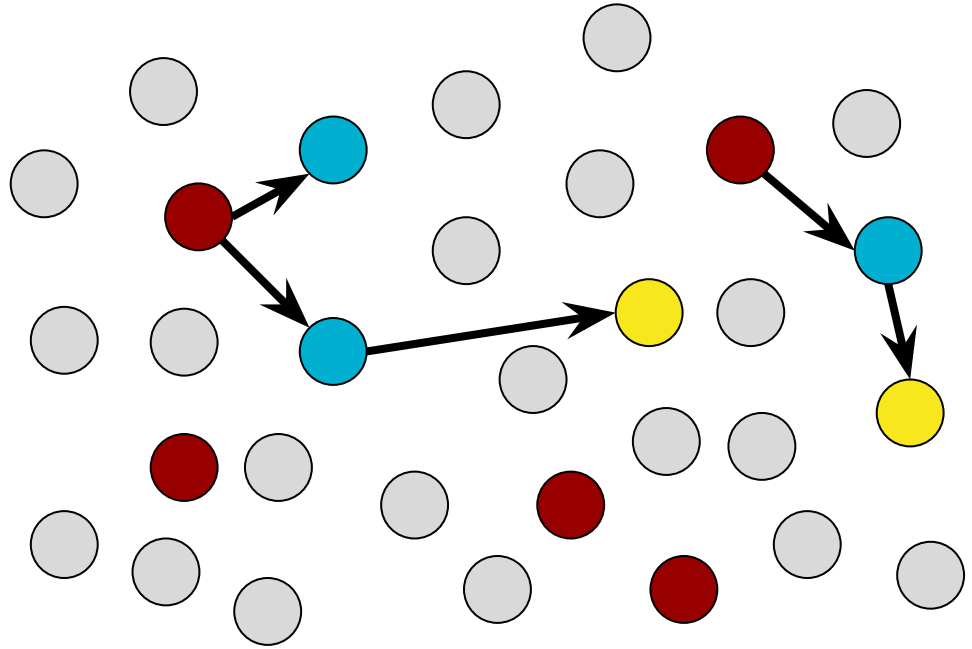


BAD!

TALEB
HALAL

WHAT IS HPSSD? How the **Violet Curve** was generated?

Sampling a **random sample**, then ask to the sampled people to share the survey questionnaire to their social contacts, link-tracing who recruited who.



INDEX

- BACKGROUND ON SOCIAL NETWORKS OF SMOKERS
- CLIQUES-AND-BLOCKS: A WAY TO BUILD A NETWORK WITH ATTRIBUTE-ASSORTATIVITY
- HOW TO MODEL A HYBRID PROBABILISTIC-SNOWBALL SAMPLING DESIGN
- PERFORMANCES AND FINAL CONSIDERATIONS

BACKGROUND: WESTERN SOCIETIES

- **Smokers are overall less connected than non-Smokers**
- **Assortativity: smokers ends to be connected a bit more with smokers, non-smokers with non-smokers**
- **Smokers are sometimes a bit isolated. In the past, it was not so (Christakis and Fowler 2008).**
- **Works for the West: in other societies, smoking is prevalent and smokers are not isolated.**

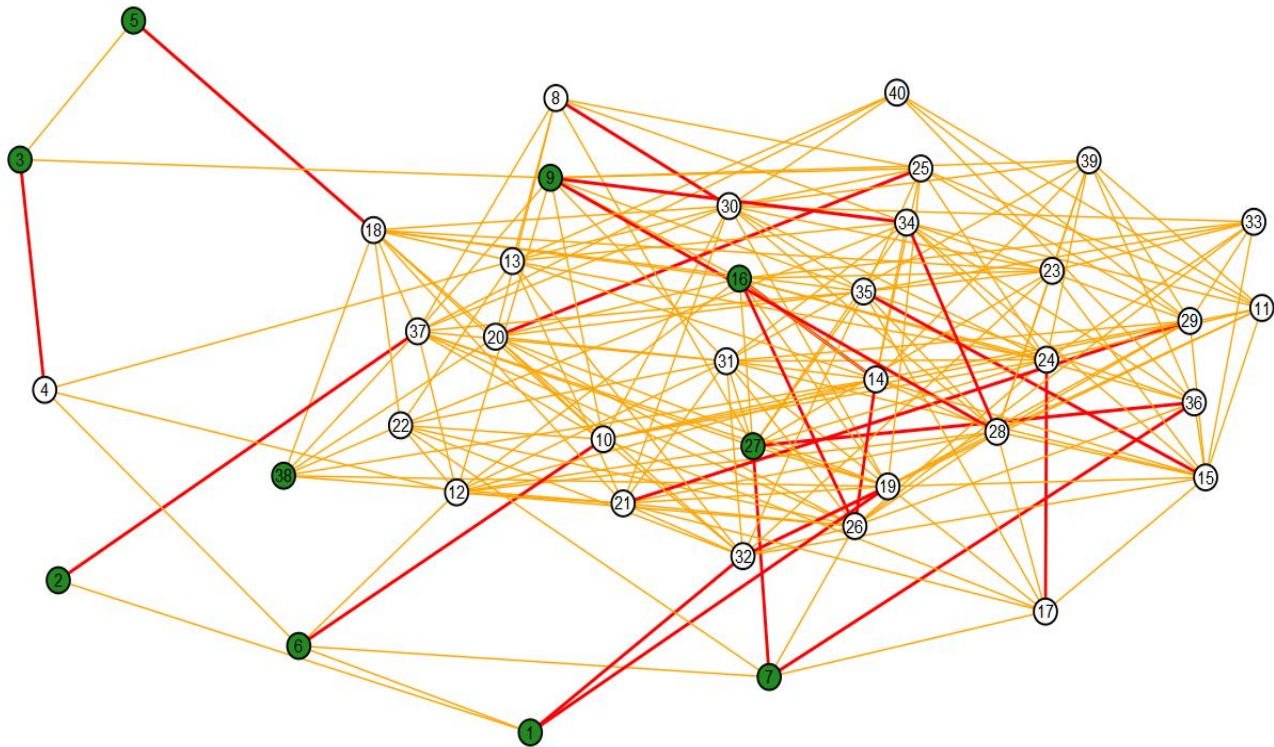
CLIQUEES AND BLOCKS, A METHOD TO GENERATE ATTRIBUTE-ASSORTATIVE NETWORKS

THE THEORY IN FIVE STEPS:

- Generate unconnected nodes
- Random connect the nodes into cliques; each clique has a propensity coefficient α for the attribute
- Left-join to the edgelist of a stochastic blockmodel, parameterised on propensity coefficient β as blocks
- Set a $e_i = f(\alpha, \beta)$ for node-level propensity scores for being a smoker
- RNG the binary attribute (smoking): nodes-per-nodes, parameterised on e_i

CLIQUEES AND BLOCKS, RESULT (EXTREME ONE)

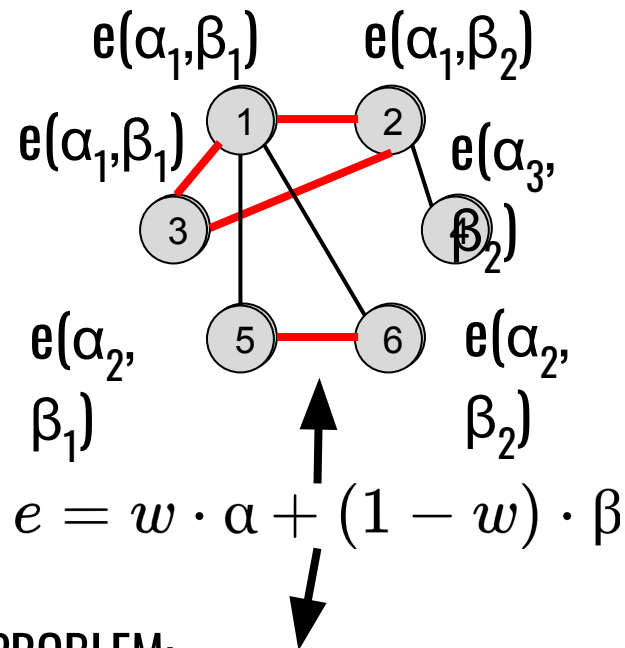
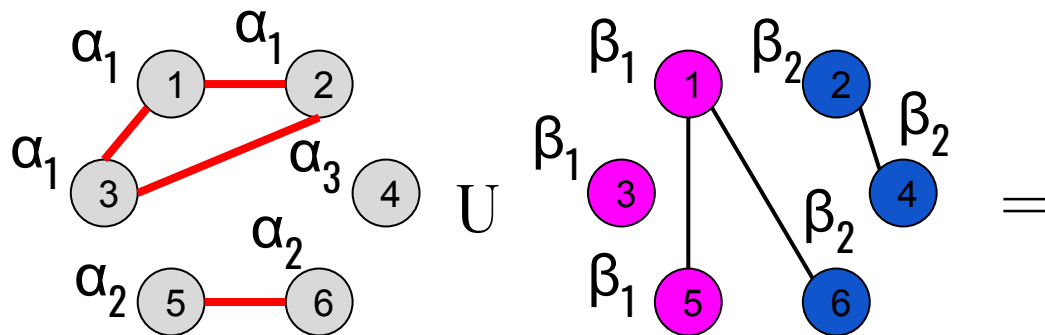
This is a case with a very high isolation of the greens and high degree homophily



GREEN NODES:
SMOKERS

RED TIES:
CLIQUEES

CLIQUEES AND BLOCKS, METHOD



“e”: probability to be a smoker, given latent “factors of smoking”. Hence: a **propensity score** for smoking.

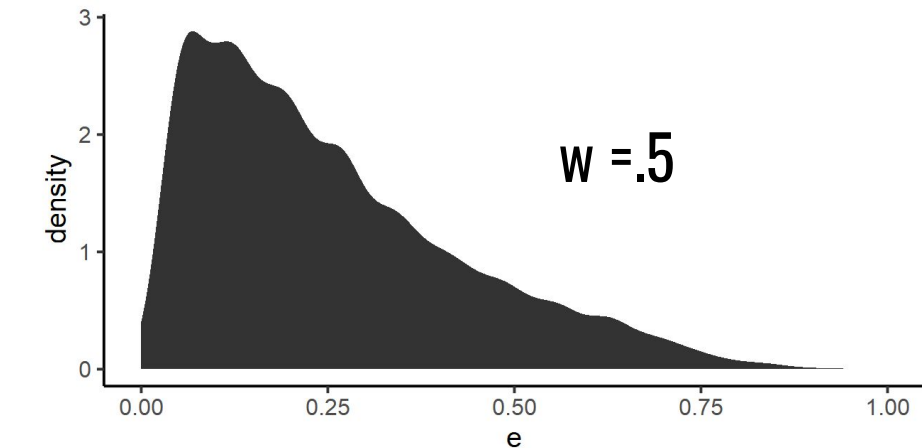
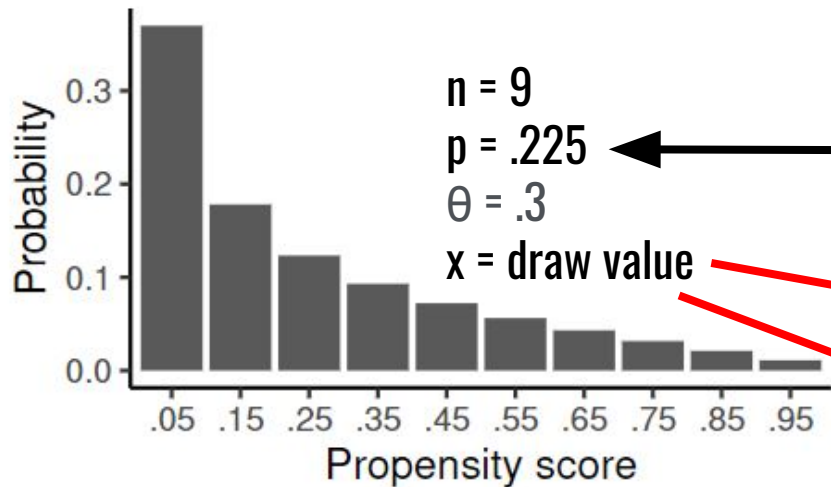
It is an abstraction, but it has a technical problem...

PROBLEM:

IF α and β are drawn from the same X ,
THEN $\text{var}(e) < \text{var}(X)$

SOLUTION: OVERDISPERSING BINOMIALS WITH A BETA-BINOMIAL MODEL

Beta-Binomial: $f(x, n, p, \theta)$



p is random,
others are fixed

$$w \cdot x_1 + (1 - w) \cdot x_2 = e$$

How, and how much, are blocks connected?

0.95	0	0	0.001	0.002	0.004	0.007	0.01	0.012	0.013	0.013
0.85	0	0.001	0.002	0.004	0.006	0.01	0.012	0.014	0.014	0.013
0.75	0	0.001	0.003	0.006	0.01	0.013	0.015	0.015	0.014	0.012
0.65	0.001	0.002	0.006	0.01	0.014	0.016	0.016	0.015	0.012	0.01
0.55	0.002	0.005	0.009	0.014	0.017	0.017	0.016	0.013	0.01	0.007
0.45	0.004	0.009	0.015	0.019	0.019	0.017	0.014	0.01	0.006	0.004
0.35	0.008	0.016	0.021	0.021	0.019	0.014	0.01	0.006	0.004	0.002
0.25	0.016	0.024	0.024	0.021	0.015	0.009	0.006	0.003	0.002	0.001
0.15	0.029	0.029	0.024	0.016	0.009	0.005	0.002	0.001	0.001	0
0.05	0.039	0.029	0.016	0.008	0.004	0.002	0.001	0	0	0
	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

Rigid smokers homophily structure

Flexible degree homophily structure

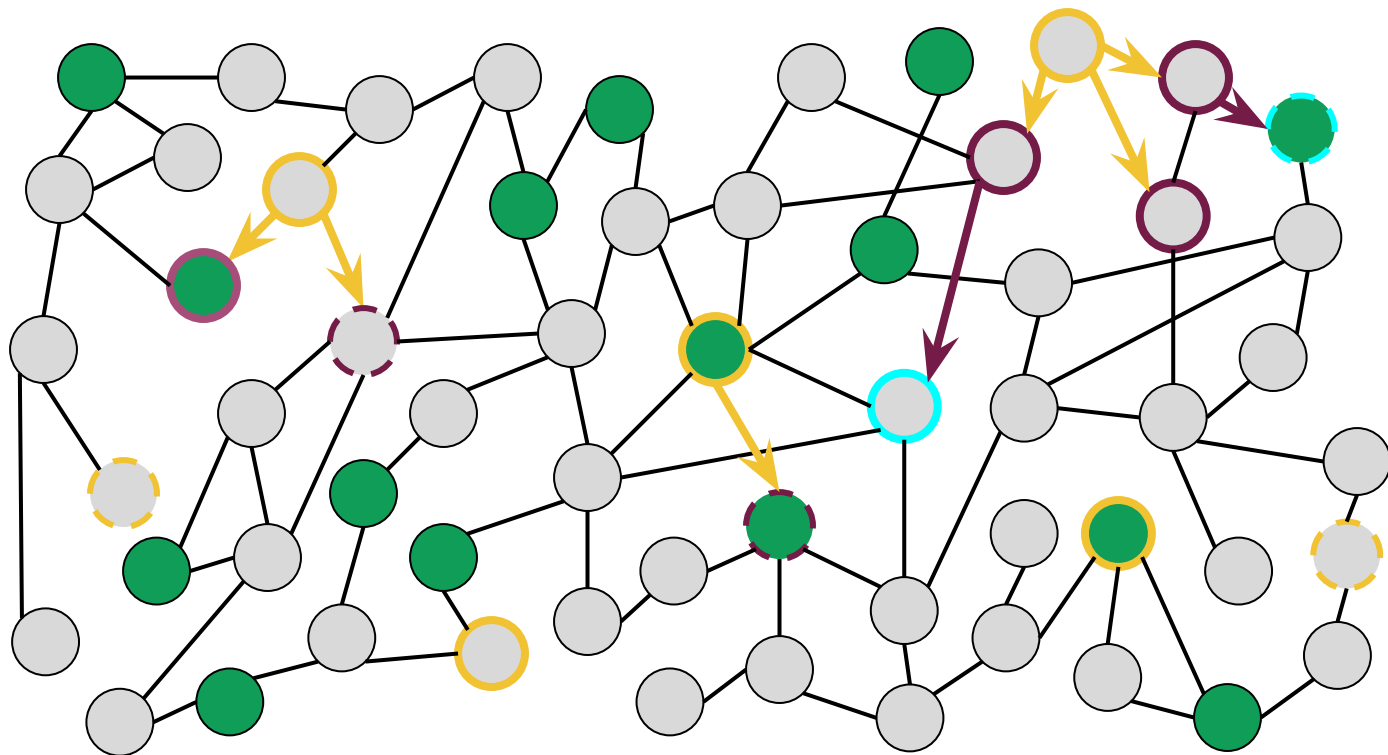
$$\pi_{\gamma} = \frac{\pi_B^{\gamma}}{\sum \pi_B^{\gamma}}$$

The higher Gamma,
the higher degree homophily

How is π_B determined?

HOW TO MODEL THE SNOWBALL RECRUITMENT?

ATTRITION IS THE BIG BAD

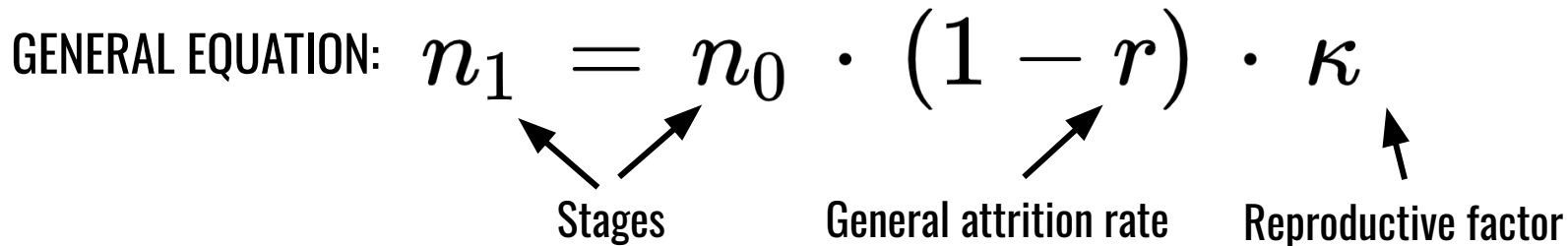


- Unrecruited Smoker
- Stage 0 Non-smoker
- Stage 0 - Refuted Non-smoker
- 1st Stage Non-smoker
- 1° Stage - Refuted Non-smoker
- n° Stage - Refuted Smoker

THERE ARE DIFFERENT SCENARIOS OF RECRUITMENT

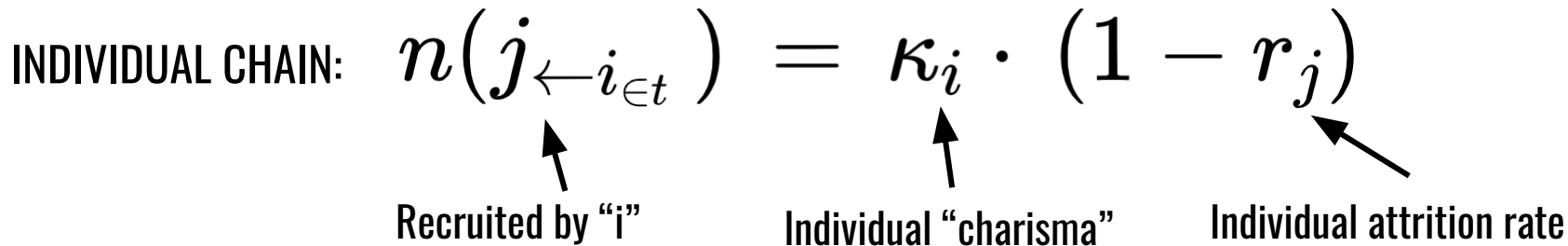
GENERAL EQUATION: $n_1 = n_0 \cdot (1 - r) \cdot \kappa$

Stages General attrition rate Reproductive factor

The diagram shows the general equation for recruitment: $n_1 = n_0 \cdot (1 - r) \cdot \kappa$. Below the equation, three labels are positioned with arrows pointing to specific parts of the equation: 'Stages' points to n_1 and n_0 ; 'General attrition rate' points to r ; and 'Reproductive factor' points to κ .

INDIVIDUAL CHAIN: $n(j_{\leftarrow i \in t}) = \kappa_i \cdot (1 - r_j)$

Recruited by "i" Individual "charisma" Individual attrition rate

The diagram shows the individual chain equation for recruitment: $n(j_{\leftarrow i \in t}) = \kappa_i \cdot (1 - r_j)$. Below the equation, three labels are positioned with arrows pointing to specific parts of the equation: 'Recruited by "i"' points to $j_{\leftarrow i \in t}$; 'Individual "charisma"' points to κ_i ; and 'Individual attrition rate' points to r_j .

THE SCALE-FREE FOREST SCENARIO (JUST ONE AMONG MANY)

THE SHIFTED YULE-SIMON MODEL

$$PMF(\kappa_i, \lambda) = \frac{\lambda \cdot (\lambda! \kappa!)}{(\lambda + \kappa + 1)!}$$

$$\mathbb{E}(\kappa_i \mid \lambda) = \frac{\lambda}{\lambda - 1} - 1$$

$$\lim_{t \rightarrow \infty} \sum_{t \in \mathbb{N}^0} [\kappa_1 \mid \lambda = 3]^t = 1 \quad \therefore \quad \forall \lambda > 3, \quad \sum_1^t n \neq \infty$$



EVALUATION: ESTIMATORS

Population Prevalence: $\vartheta_{design}(y) = \bar{y}$

THIS NEED TO BE FURTHER DISCUSSED AT THE END

Bias: Sampling Error

$$SE_{\nu|design} = \vartheta(y_{\nu}) - y_{\nu}$$
$$bias = \langle SE_{design} \rangle$$

EVALUATION: ESTIMATORS

Precision: Design Effect

$$DE = \frac{\sigma^2(\vartheta_1(y))}{\sigma^2(\vartheta_0(y))}$$

Here ϑ_0 is the **Golden Design**

Performance:

- Absolute Error in the Sample
- Expected Impact
After Switching Design

$$AE_{\vartheta} = |y - \vartheta(y)|$$

$$\mathbb{E}(I_{\vartheta_0 \rightarrow 1}) = \left\langle \frac{AE_{\vartheta_0}}{y} - \frac{AE_{\vartheta_1}}{y} \right\rangle$$

Here ϑ_0 can be any alternative

INTERPRETATION OF “EXPECTED IMPACT AFTER SWITCHING”

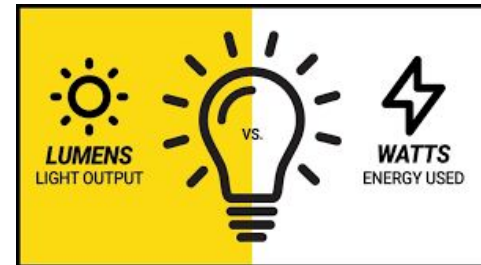
It is how better-or-worse the “cheap” design will perform compared to the expensive alternative. **This could also be computed non-parametrically with minor fixes.**

So, it makes sense to implement it into an efficiency function: $\frac{\mathbb{E}(I)}{n_0}$

Statistics are notoriously hard to generalize, the function would not be analytically derivable - to my knowledge it can only be empirically inferred.

Generally, heuristics are followed:

- if you can reach $n_0 = 1000$, you will be fine no matter what
- if you cannot, **estimate a Confidence Interval for errors (or any other Bayesian diablerie like Credible Interval).** **Actually not covered in slides. Covered in paper.**



4 SCENARIOS: EACH ITERATION FOLLOWS ALL OF THEM

- Same cost than Gold Standard, Poisson($\lambda=.5$) recruitment
- Same cost than Gold Standard, Yule($\lambda=.3$) recruitment
- Half cost than Gold Standard, Poisson($\lambda=.5$) recruitment
- Half cost than Gold Standard, Yule($\lambda=3$) recruitment

6 HYPER-PARAMETERS: RANDOM ACROSS ITERATIONS

- N. of Cliques
- $\langle y \rangle$: Prevalence
- γ : Homophily
- Neighborhood size
- w : Familism
- r : Attrition

Preliminary Results: COEFFICIENTS IN SCENARIO I

term	estimate	std.error	p.value
abs_gold	0.373	0.010	0.000
gamma	0.207	0.011	0.000
phi_y	0.193	0.011	0.000
phi_k	0.189	0.011	0.000
Q	0.164	0.011	0.000
w	-0.112	0.011	0.000
r	-0.093	0.011	0.000
k	0.082	0.011	0.000
y	0.053	0.011	0.000
N	0.030	0.011	0.007

**NOTE:
THIS IS FOR
COMPARATIVE
INFERENCES,
NOT PREDICTIONS,
SINCE TERMS ARE
STANDARDISED**

Preliminary Results: COEFFICIENTS IN SCENARIO II

term	estimate	std.error	p.value
abs_gold	0.237	0.011	0.000
phi_y	0.157	0.011	0.000
gamma	0.146	0.011	0.000
Q	0.139	0.011	0.000
phi_k	0.136	0.011	0.000
w	-0.090	0.011	0.000
y	0.054	0.011	0.000
k	0.051	0.011	0.000
r	-0.045	0.011	0.000
N	0.014	0.011	0.197

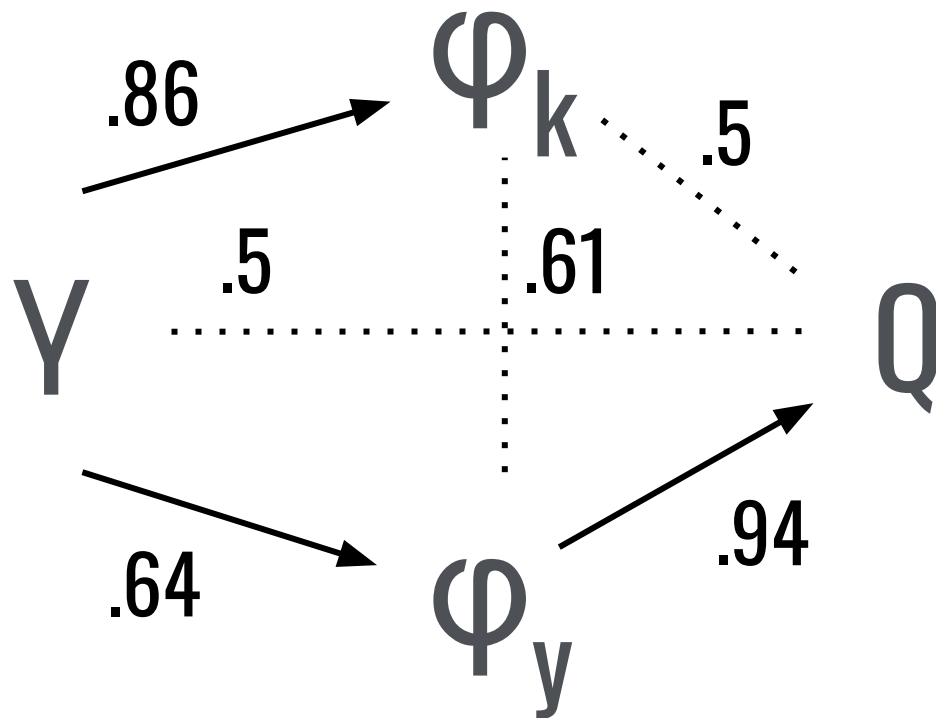
The interpretation
of abs_gold is very
relevant:

Randomicity within
and across Monte
Carlo universes

Preliminary Results: COEFFICIENTS IN SCENARIO III

term	estimate	std.error	p.value
abs_gold	0.379	0.010	0.000
gamma	0.211	0.011	0.000
phi_y	0.199	0.011	0.000
phi_k	0.190	0.011	0.000
Q	0.174	0.011	0.000
w	-0.108	0.011	0.000
k	0.086	0.011	0.000
r	-0.085	0.011	0.000
y	0.065	0.011	0.000
N	0.035	0.011	0.002

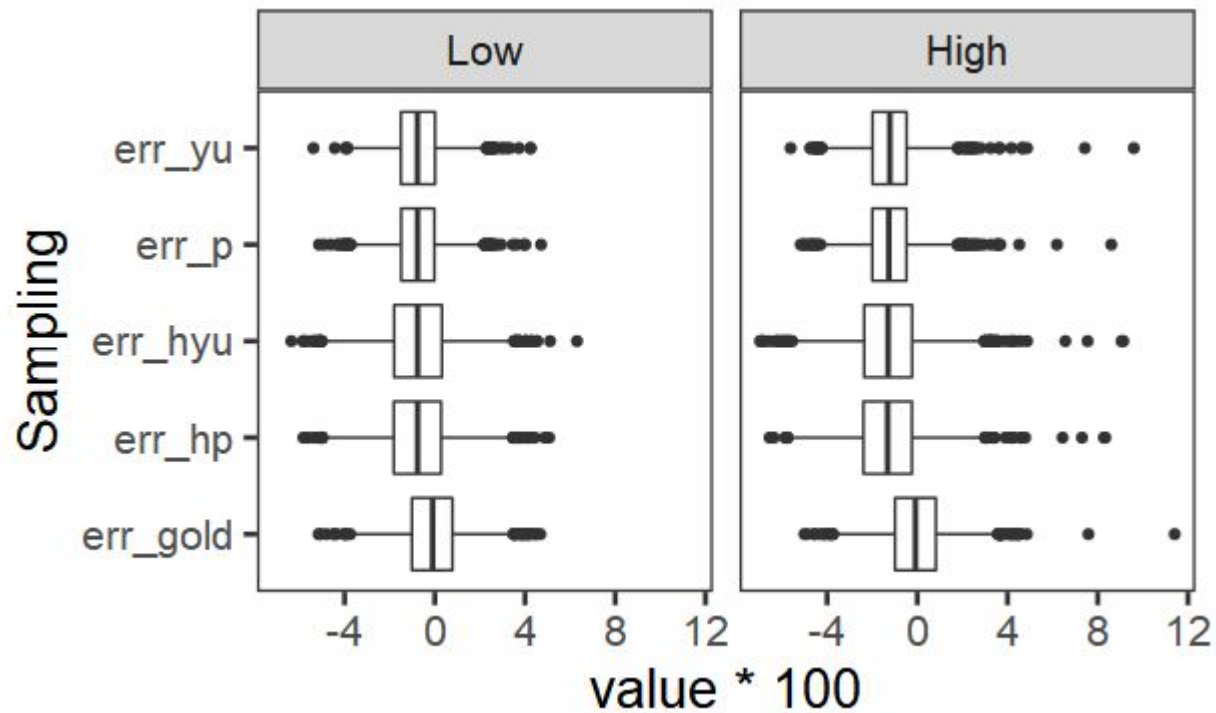
GAMMA IS THE BEST PREDICTOR OF PERFORMANCE IN HPSS



Y	ϕ_y	ϕ_k	Q
Low	.017	.047	.006
High	.025	.11	.008

Modularity is not very present.

BIAS AND PRECISION OF HPSSD



SO, IS IT WORTH THIS HPSSD?

GAMMA

	Low <dbl>	Mid-Low <dbl>	Mid-High <dbl>	High <dbl>
Scenario I	0.251	-0.368	-0.874	-1.732
Scenario II	-1.039	-1.529	-2.037	-2.736
Scenario II*	0.760	0.502	-0.290	-0.927
Scenario III	0.247	-0.360	-0.799	-1.735
Scenario IV	-1.035	-1.588	-1.919	-2.728
Scenario IV*	0.764	0.443	-0.172	-0.919

Impacts are multiplied per 100.
They are multipliers of
“efficiency”.

Final heuristic: if you cannot reach $n = 1000$,

- link-trace
- calculate homophily in link-traced chain
- if **not sign**, then count in the snowball component

LET'S SHOW SOME NON-PARAMETRIC STATISTICS

Frequency of HPSSD being better than alternative

GAMMA

	Low <dbl>	Mid-Low <dbl>	Mid-High <dbl>	High <dbl>
Scenario I	0.52	0.44	0.38	0.29
Scenario II	0.42	0.39	0.35	0.31
Scenario II*	0.55	0.54	0.45	0.40
Scenario III	0.52	0.44	0.39	0.31
Scenario IV	0.42	0.38	0.35	0.31
Scenario IV*	0.55	0.54	0.47	0.40

THE MOST IMPORTANT FINAL DISCUSSION

When we link trace, we collect 3 new information:

- Differences between recruiter and recruited
- Chains of recruitment
- Size of the chains

With these infos we can try to estimate homophily in the population.

Could we use these to de-bias HPSSD?

I tried and I failed.

THE END.