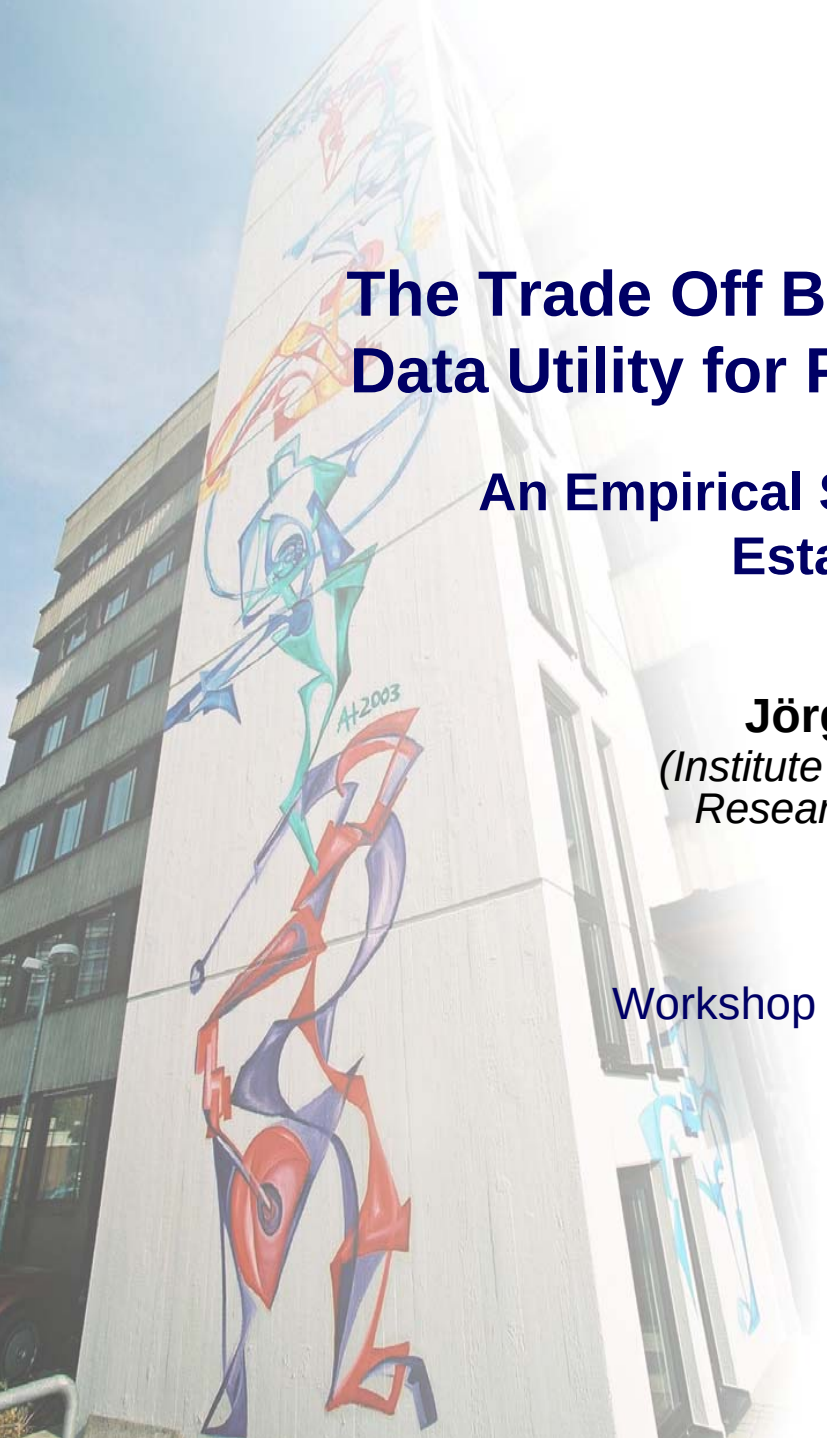


The Trade Off Between Disclosure Risk and Data Utility for Partially Synthetic Data Sets

An Empirical Study using the German IAB Establishment Survey

Jörg Drechsler *(Institute for Employment Research, Germany)*

&

Jerry Reiter *(Duke University)*

Workshop on Data Access to Micro-Data, Nürnberg

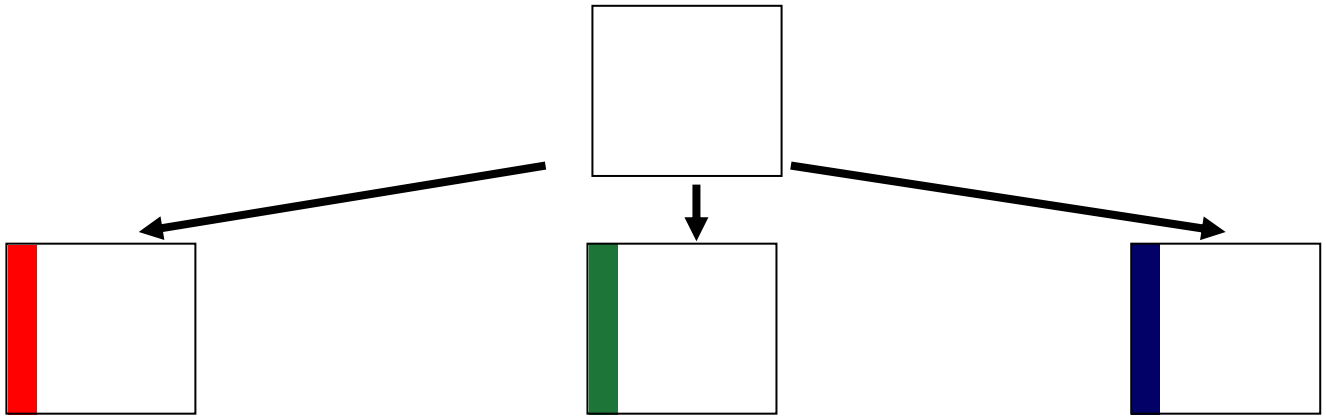
20.-21. August 2007

Overview

- Partially Synthetic Data Sets for Statistical Disclosure Control
- Partially Synthetic Data Sets for the IAB Establishment Panel
- Data Utility
- Disclosure Risk
- Two Stage Imputation
- Conclusion/Future Work

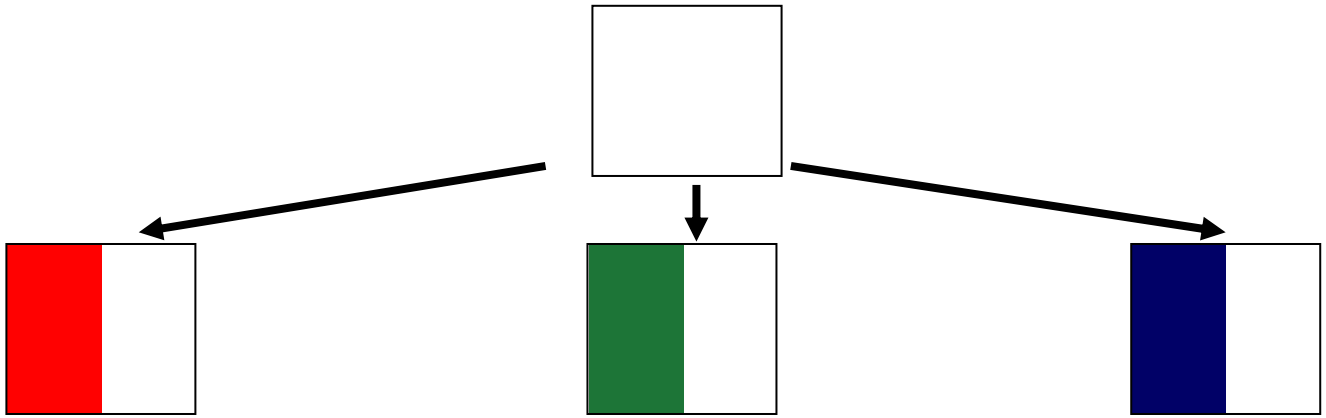
Partially Synthetic Data Sets

- values of key identifiers or sensitive variables are replaced by synthetic values



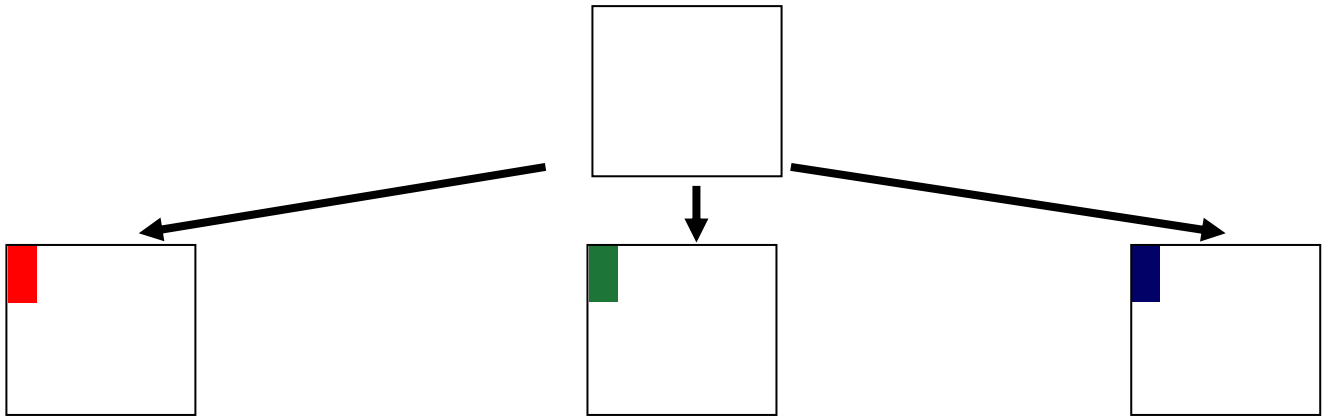
Partially Synthetic Data Sets

- values of key identifiers or sensitive variables are replaced by synthetic values



Partially Synthetic Data Sets

- values of key identifiers or sensitive variables are replaced by synthetic values



Overview

- Partially Synthetic Data Sets for Statistical Disclosure Control
- Partially Synthetic Data Sets for the IAB Establishment Panel
- Data Utility
- Disclosure Risk
- Two Stage Imputation
- Conclusion/Future Work

The IAB Establishment Panel

- Annually conducted Establishment Survey
- Since 1993 in Western Germany, since 1996 in Eastern Germany
- *Population*: All establishments with at least one employee covered by social security
- *Source*: Official Employment Statistics
- Response rate of repeatedly interviewed establishments more than 80%
- Sample of more than 16.000 establishments in the last wave
- *Contents*: employment structure, changes in employment, business policies, investment, training, remuneration, working hours, collective wage agreements, works councils

Imputation Procedure

- Only two variables are synthesized: - number of employees
  - industry (16 categories)
- Imputation by sequential regression
- Imputation model: - linear model for the cubic root of the nb of employees
  - multinomial logit for the industry
- No interactions
- No missing values in original data
- Additional information for the imputation model from the German Social Security Data
- Different number of imputations to evaluate the impact on disclosure risk and data utility (*m=3,10,50,100*)

Overview

- Partially Synthetic Data Sets for Statistical Disclosure Control
- Partially Synthetic Data Sets for the IAB Establishment Panel
- Data Utility
- Disclosure Risk
- Two Stage Imputation
- Conclusion/Future Work

Analytical Validity

- Compare regression results from the original data with results from the synthetic data
- Zwick (2005) analyses the productivity effects of different continuing vocational training forms in Germany
- Probit regression to explain, why firms offer vocational training
- 13 Explanatory variables including: Share of qualified employees, establishment size, industry, collective wage agreement, high qualification needs expected…

- Two Data Utility Measures:

  - relative deviation between beta coefficients from the original data set and the synthetic data sets
  - confidence interval overlap

Estimates from the regression

exogenous variables	*Coef. Org*	*m=3*	*m=10*	*m=50*	*m=100*
Redundancies expected	0.253	0.283	0.276	0.277	0.2759
Many emp. exp. to be on maternity leave	0.262	0.342	0.337	0.328	0.3290
High qualification need exp.	0.646	0.639	0.645	0.648	0.6506
Appr. training reaction on skill shortages	0.113	0.122	0.103	0.098	0.0978
Training reaction on skill shortages	0.540	0.544	0.533	0.531	0.5300
Establishment size 20-199	0.684	0.477	0.478	0.475	0.4756
Establishment size 200-499	1.352	1.151	1.190	1.182	1.1818
Establishment size 500-999	1.346	1.467	1.442	1.417	1.4187
Establishment size 1000 +	1.955	1.979	2.065	2.162	2.2031
Share of qualified employees	0.787	0.825	0.823	0.827	0.8225
State-of-the-art technical equipment	0.171	0.177	0.172	0.170	0.1696
Collective wage agreement	0.255	0.330	0.320	0.305	0.3061
Apprenticeship training	0.490	0.546	0.544	0.541	0.5436
Eastern Germany dummy	0.058	0.074	0.070	0.072	0.0719
industry dummy 1	-0.793	-0.635	-0.694	-0.723	-0.7285
industry dummy 2	0.120	-0.160	0.025	0.074	0.0963
industry dummy 3	-0.264	-0.359	-0.298	-0.253	-0.2754
industry dummy 4	-0.196	-0.137	-0.165	-0.170	-0.1708
industry dummy 5	-0.621	-0.660	-0.665	-0.621	-0.6382
industry dummy 6	-0.658	-0.651	-0.642	-0.633	-0.6356
industry dummy 7	-0.191	-0.249	-0.243	-0.221	-0.2285
industry dummy 9	0.771	0.736	0.605	0.631	0.6378
industry dummy 10	0.202	-0.068	0.051	0.032	0.0487
industry dummy 11	-0.349	-0.292	-0.314	-0.319	-0.3268
industry dummy 12	0.158	0.343	0.201	0.197	0.1946
industry dummy 13	0.318	0.341	0.280	0.289	0.2930
industry dummy 14	0.075	0.038	0.069	0.031	0.0333
constant	-1.319	-1.353	-1.325	-1.316	-1.3113

Relative deviation between synthetic and original estimates

exogenous variables	*m=3*	*m=10*	*m=50*	*m=100*
Redundancies expected	11.75%	8.77%	9.43%	8.91%
Many emp. exp. to be on maternity leave	30.42%	28.48%	24.88%	25.42%
High qualification need exp.	-1.03%	-0.09%	0.27%	0.71%
Appr. training reaction on skill shortages	8.41%	-8.55%	-12.98%	-13.34%
Training reaction on skill shortages	0.87%	-1.18%	-1.56%	-1.79%
Establishment size 20-199	-30.24%	-30.19%	-30.62%	-30.51%
Establishment size 200-499	-14.87%	-12.00%	-12.57%	-12.61%
Establishment size 500-999	9.00%	7.10%	5.28%	5.38%
Establishment size 1000 +	1.20%	5.60%	10.56%	12.68%
Share of qualified employees	4.81%	4.58%	5.11%	4.49%
State-of-the-art technical equipment	3.70%	0.55%	-0.52%	-0.69%
Collective wage agreement	29.10%	25.24%	19.51%	19.91%
Apprenticeship training	11.32%	10.99%	10.26%	10.83%
Eastern Germany dummy	27.55%	21.44%	24.80%	24.11%
industry dummy 1	-19.96%	-12.43%	-8.81%	-8.13%
industry dummy 2	-233.99%	-79.44%	-38.14%	-19.42%
industry dummy 3	35.93%	12.86%	-4.26%	4.35%
industry dummy 4	-30.31%	-15.79%	-13.29%	-13.05%
industry dummy 5	6.33%	7.16%	0.12%	2.82%
industry dummy 6	-1.02%	-2.33%	-3.76%	-3.33%
industry dummy 7	30.46%	27.23%	15.41%	19.51%
industry dummy 9	-4.47%	-21.52%	-18.08%	-17.24%
industry dummy 10	-133.73%	-74.86%	-84.16%	-75.87%
industry dummy 11	-16.29%	-9.87%	-8.47%	-6.29%
industry dummy 12	117.38%	27.56%	24.95%	23.40%
industry dummy 13	7.16%	-12.04%	-9.12%	-7.94%
industry dummy 14	-48.52%	-7.19%	-58.22%	-55.46%
constant	2.58%	0.48%	-0.18%	-0.55%

Comparison for the deviations

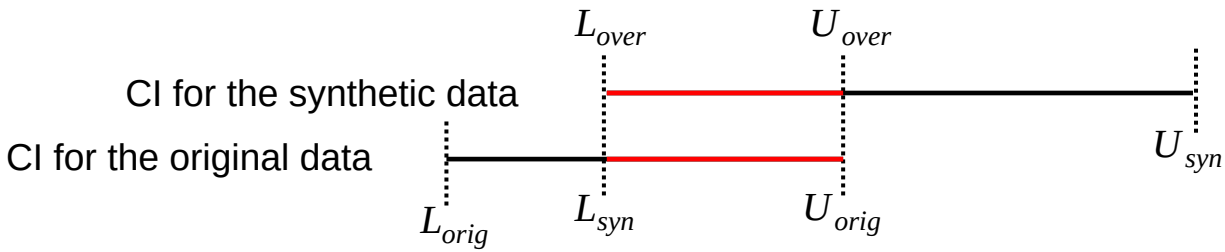
- Total number of estimates: 28

deviation between original and synthetic value	*m=3 to m=10*	*m=10 to m=50*	*m=50 to m=100*
better	21	16	14
worse	7	12	14

Confidence Interval Overlap

- Suggested by Karr et al. (2006)
- Measure the overlap of CIs from the original data and CIs from the synthetic data
- The higher the overlap, the higher the data utility
- Compute the average relative CI overlap for any β_k

$$J_k = \frac{1}{2}\left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right]$$



Confidence Interval Overlap

exogenous variables	*m=3*	*m=10*	*m=50*	*m=100*
Redundancies expected	0.858	0.894	0.886	0.892
Many emp. exp. to be on maternity leave	0.789	0.801	0.826	0.822
High qualification need exp.	0.976	0.998	0.994	0.984
Appr. training reaction on skill shortages	0.953	0.953	0.928	0.926
Training reaction on skill shortages	0.975	0.967	0.956	0.949
Establishment size 20-199	0.000	0.000	0.000	0.000
Establishment size 200-499	0.326	0.459	0.429	0.427
Establishment size 500-999	0.737	0.793	0.845	0.842
Establishment size 1000 +	0.927	0.864	0.519	0.513
Share of qualified employees	0.862	0.867	0.852	0.869
State-of-the-art technical equipment	0.960	0.994	0.994	0.992
Collective wage agreement	0.571	0.624	0.709	0.702
Apprenticeship training	0.669	0.679	0.699	0.682
Eastern Germany dummy	0.899	0.922	0.909	0.912
industry dummy 1	0.564	0.723	0.802	0.816
industry dummy 2	0.649	0.883	0.943	0.961
industry dummy 3	0.806	0.911	0.969	0.969
industry dummy 4	0.786	0.903	0.912	0.914
industry dummy 5	0.893	0.863	0.998	0.946
industry dummy 6	0.960	0.947	0.913	0.923
industry dummy 7	0.804	0.816	0.894	0.866
industry dummy 9	0.800	0.823	0.850	0.856
industry dummy 10	0.671	0.819	0.795	0.816
industry dummy 11	0.840	0.896	0.908	0.932
industry dummy 12	0.545	0.888	0.894	0.900
industry dummy 13	0.946	0.904	0.926	0.935
industry dummy 14	0.902	0.981	0.872	0.878
constant	0.901	0.981	0.993	0.978

Comparison of the overlap

CI overlap	m=3 to m=10	m=10 to m=50	m=50 to m=100
better	22	17	14
worse	6	11	14

Average overlap

$$J = (1/p)\sum_{i=1}^{p} J_k$$

	m=3	m=10	m=50	m=100
Average CI overlap	0.7703	0.8269	0.8291	0.8286

Overview

- Partially Synthetic Data Sets for Statistical Disclosure Control
- Partially Synthetic Data Sets for the IAB Establishment Panel
- Data Utility
- Disclosure Risk
- Two Stage Imputation
- Conclusion/Future Work

Disclosure Risk

- 3 different disclosure risk measures by Reiter & Mitra (to appear)
- Assumptions about the intruder:
  - Intruder knows which establishments participated in the survey
  - Intruder has exact information on the number of employees and the industry code for these establishments (target records)
- Observation from the survey is considered a match if

$$industry_{\text{target}} = industry_{obs} \quad \& \quad nb.emp_{\text{target}} - sd_s(nb.emp_{\text{target}}) \leq nb.emp_{obs} \leq nb.emp_{\text{target}} + sd_s(nb.emp_{\text{target}})$$

$sd_s(nb.emp_{obs})$ standard dev. for the nb of emp. in cell *s*

- 150 different cells *s* defined by quantiles

Perceived match risk

- Number of target records with a maximum match probability exceeding some predefined threshold
- Match probability $p_m = 1/n_m$

n_m number of observations that fulfil matching conditions

Example:

	number of obs fulfilling matching conditions		
target record	synthetic data1	synthetic data2	synthetic data3
1	12	8	11
2	3	5	7
3	9	4	15
n	32	43	36

	match probability			
target record	synthetic data1	synthetic data2	synthetic data3	max
1	1/12	1/8	1/11	0.13
2	1/3	1/5	1/7	0.33
3	1/9	1/4	1/15	0.25
n	1/32	1/43	1/36	0.03

Expected match risk

$$p_{\exp} = \sum_j (1/c_j) I_j$$

$$c_j = \min(n_m^{(i)})$$

$$I_j = \begin{cases} 1 & \text{if true match is among } c_j \\ 0 & \text{otherwise} \end{cases}$$

target record	number of obs fulfilling matching conditions synthetic data1	synthetic data2	synthetic data3	I	$(1/c_j)I_j$
1	12	8	11	1	1/8
2	3	5	7	0	0
3	9	4	15	1	1/4
					0
n	32	43	36	1	1/32

True match risk

$$p_{true} = \sum_j K_j$$

$$K_j = \begin{cases} 1 & \text{if } c_j I_j = 1 \\ 0 & \text{otherwise} \end{cases}$$

Disclosure risk for different m

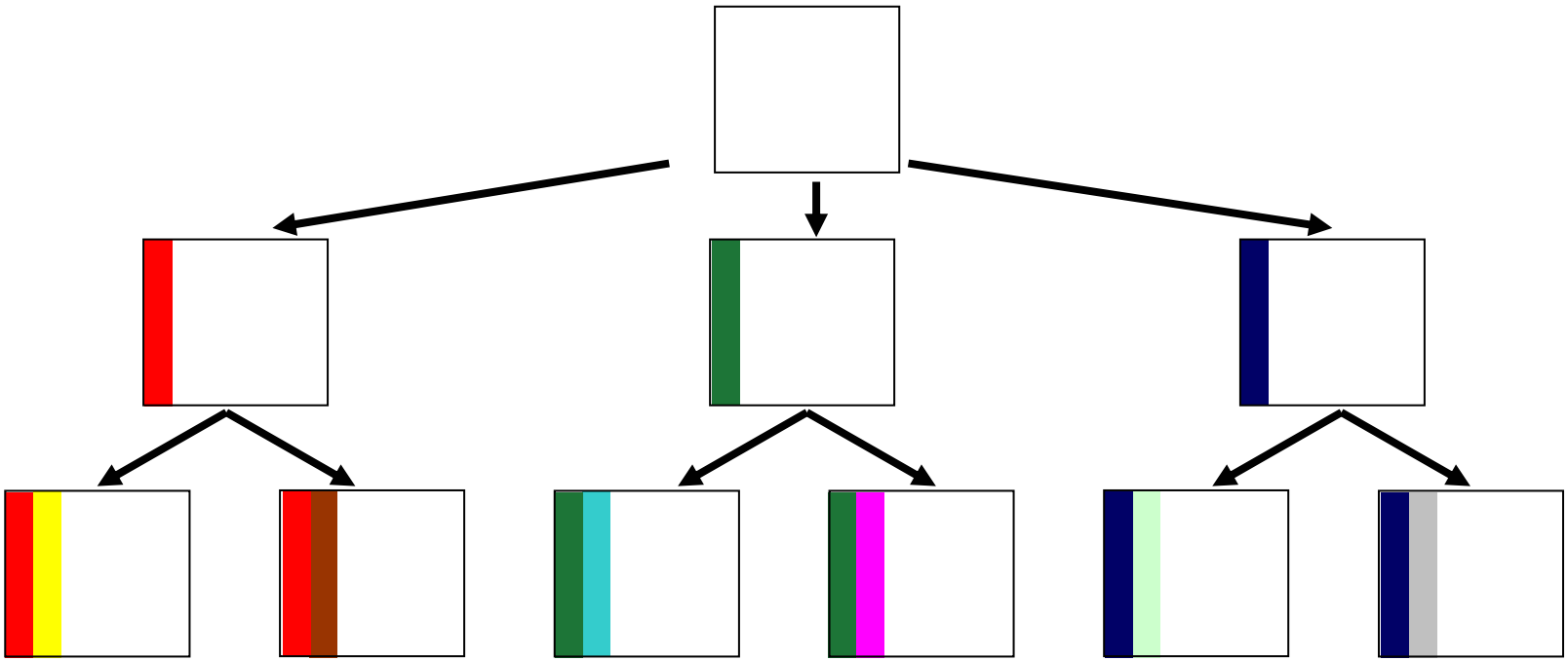
	m=3	*m=10*	*m=50*	*m=100*
perceived match risk	3649	3806	3818	3819
expected match risk	36.4931	40.5332	40.6252	39.7862
true match risk	19	30	35	35

Overview

- Partially Synthetic Data Sets for Statistical Disclosure Control
- Partially Synthetic Data Sets for the IAB Establishment Panel
- Data Utility
- Disclosure Risk
- Two Stage Imputation
- Conclusion/Future Work

Multiple Imputation in Two Stages

- some variables are imputed more often than others



Different two stage settings

- First stage: industry code is imputed *m* times
- Second stage: number of employees is imputed *r* times within each first stage nest
- 6 different settings:
  - m=3, r=3
  - m=3, r=16
  - m=3, r=33
  - m=5, r=10
  - m=16, r=3
  - m=10, r=5

Data Utility & Disclosure Risk for m=3

	m=3		*r=3 to r=1*		*m,r=3 to m=10*	
	better	*worse*	*better*	*worse*	*better*	*worse*
dev. betw. org. and syn. value	-	-	20	8	11	17
CI overlap	-	-	18	10	9	19
Average CI overlap	0.7703		0.7965		0.8269	

m=3	*r=1*	*r=3*	*m=10*
perceived match risk	3649	3791	3806
expected match risk	36.493	37.209	40.533
true match risk	19	26	30

DU & DR for 50 synthetic data sets

	m=50		m=3,r=16 to m=50		m=5,r=10 to m=50		m=10 r=5 to m=50		m=16,r=3 to m=50	
	better	worse	better	worse	better	worse	better	worse	better	worse
deviation org./syn.	-	-	7	21	10	18	14	14	13	15
CI overlap	-	-	8	20	9	19	14	14	14	14
Average CI overlap	0.8291		0.7868		0.8066		0.8363		0.8247	

	m=50	m=3 r=16	m=5 r=10	m=10 r=5	m=16 r=3
perceived match risk	3818	3812	3814	3817	3818
expected match risk	40.6252	39.5324	45.9117	40.0927	42.5447
true match risk	35	32	40	33	37

Overview

- Partially Synthetic Data Sets for Statistical Disclosure Control
- Partially Synthetic Data Sets for the IAB Establishment Panel
- Data Utility
- Disclosure Risk
- Two Stage Imputation
- Conclusion/Future Work

Conclusions

- Disclosure risk and data utility increase with *m*
- Two stage imputation can help to address the trade off between disclosure risk and data utility
- Finding the best *m* and *r* is difficult

Future Work

- Impute number of employees at stage one and industry at stage two
- Think about better ways of defining bounds for the disclosure risk for continuous variables

Thank you for your attention