(Almost) All of Entity Resolution

Rebecca C. Steorts
joint work with **Olivier Binette**, PhD Student, Department of
Statistical Science Duke University

Department of Statistical Science, affiliated faculty in
Computer Science, Biostatistics and Bioinformatics, the
information initiative at Duke (iiD) and
the Social Science Research Institute (SSRI)
Duke University and U.S. Census Bureau

October 4, 2020

Entity resolution (record linkage or de-duplication) is the process of removing duplicated information from large noisy databases.

Entity resolution (record linkage or de-duplication) is the process of removing duplicated information from large noisy databases.

The purpose of this review is to introduce one to the fundamentals of entity resolution, its applications, and modern developments over the past $61+$ years.

What is the purpose of entity resolution?

Record Linkage[*]

HALBERT L. DUNN, M.D., F.A.P.H.A.

Chief, National Office of Vital Statistics, U. S. Public Health Service, Federal Security Agency, Washington, D. C.

EACH person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

The Book has many pages for some

the various important records of a person's life.

The two most important pages in the Book of Life are the first one and the last one. Consequently, in the process of record linkage the uniting of the fact-of-death with the fact-of-birth has been given a special name, "death clearance."

J. J. HENTZ.
BICYCLE.
(Application filed Nov. 3, 1897.)

(No Model.)

Fig. 3.

Fig. 1.

Fig. 4.

Fig. 5.

Fig. 2.
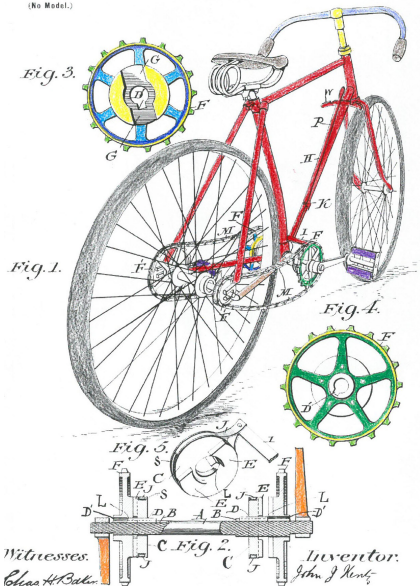
Witnesses.
Chas H. Baker

Inventor.
John J. Hentz.

r bicycle, 1899.

[An] Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

— U.S. Constitution, Article I, Section 2

By **Ted Enamorado**

Recently, there's been an uproar about Georgia's approach to voter registration. The state's "exact match" law, passed last year, requires that citizens' names on their government-issued IDs must precisely match their names as listed on the voter rolls. If the two don't match, additional verification by a local registrar will be necessary.

By **Ted Enamorado**

Oct. 20, 2018 at 5:00 a.m. MDT

Recently, there's been an uproar about Georgia's approach to voter registration. The state's "exact match" law, passed last year, requires that citizens' names on their government-issued IDs must precisely match their names as listed on the voter rolls. If the two don't match, additional verification by a local registrar will be necessary.

This is an example of a situation where a voter named "Roberto Juan Hernandez Ruiz" and "Roberto Ruiz" would be flagged.

Terminology

The linkage graph
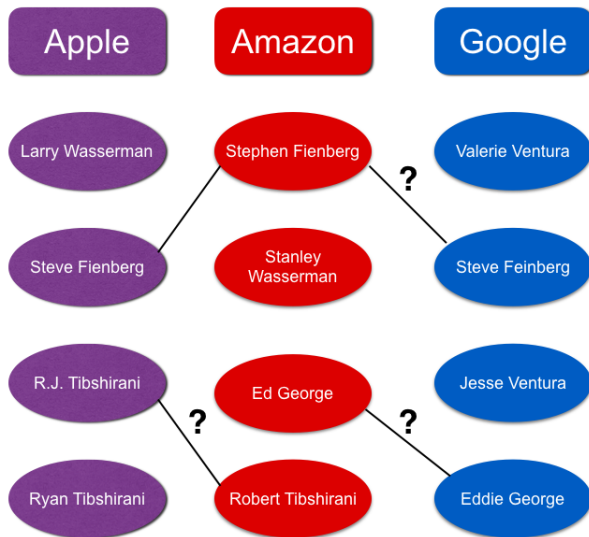
The attribute (full name) of Larry Wasserman

The collection of the record of Larry Wasserman

Larry Wasserman

1014 Murray Hill Avenue
Pittsburgh, PA 15217
412-361-3146

De-duplication, Record linkage, and Entity resolution
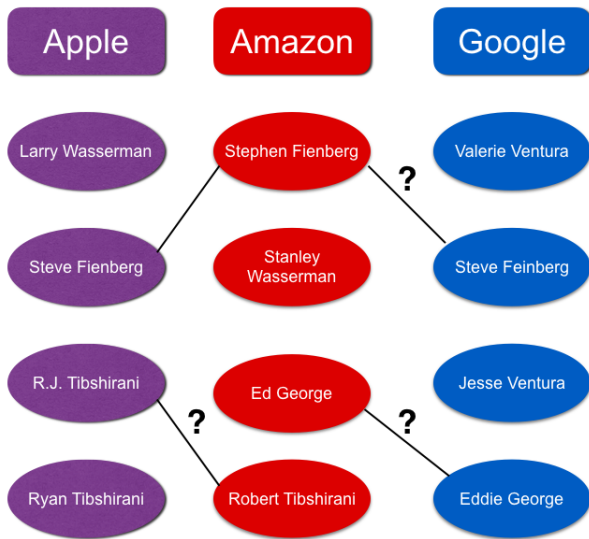
Challenges

Challenges of Entity Resolution

Costly manual labelling

Vast amounts of manually-labelled data are typically required for supervised learning and evaluation.

Scalability/computational efficiency

Approximations are required to avoid quadratic scaling. Need to ensure impact on accuracy is minimal.

Limited treatment of uncertainty

Given inherent uncertainties, it's important to output predictions with confidence regions.

Unreliable evaluation

Standard evaluation methods return imprecise estimates of performance.

Pipeline Approach

Data Cleaning Pipeline
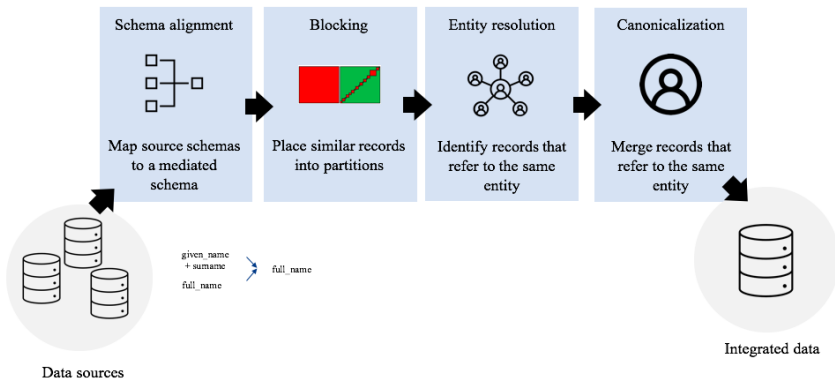
Blocking and Deterministic Record Linkage
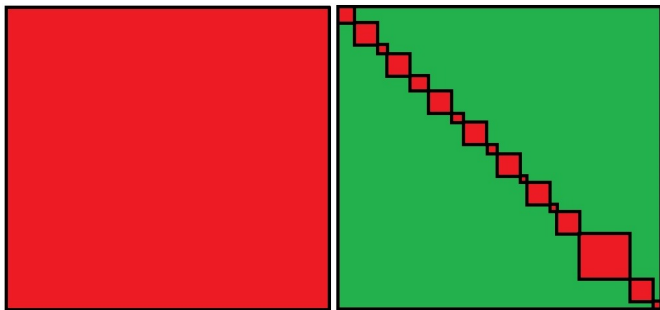
Figure: Left: All to all record comparisons. Right: Similar records placed into the same partition.

Blocking places similar records into a partition.

Deterministic record linkage is the most widely used in the literature given that it is

1. scalable
2. rules that can easily be put together
3. and it is easily transferable across disciplines

Deterministic Record Linkage

1. Exact Matching and Off-by-k Matching
2. Scoring Functions (Edit and Jaro-Winkler distances)
3. Putting simple rules together to form complex rules

Case Study from the UNTC

Case study from a human rights conflict in El Salvador from 1980-1992 using data from the United Nations Commission on the Truth (UNTC).

Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

Table: Duplicated records reproduced from Table 1 of Sadinle (2014). Records 1 – 3 should refer to the same entity. Records 4 – 5 *might* refer to the same entity. Note that record 5 most likely has OCR errors, where "RMAOS" should be "RAMOS."

Sadinle (2014) utilized complex rules for a blocking criteria.

Then he applied probabilistic record linkage within each block.

Similar rules are used in other human rights applications by Ball (2006); Sadosky, Shrivastava, Price, Steorts (2015); Chen, Shrivastava, Steorts (2018); and in work established by the Human Rights Data Analysis Group (HRDAG).

While deterministic rules are recommended for intuition or blocking, they are not recommended in place of probabilistic record linkage.

Steorts, Ventura, Sadinle, Fienberg (2014) and Murray (2016) provide reviews on deterministic and probabilistic blocking.

Probabilistic Record Linkage

Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.

*Chief, National Office of Vital Statistics, U. S. Public Health Service,
Federal Security Agency, Washington, D. C.*

Halbert L. Dunn (1896-1975):

- chief of the National Office of Vital Statistics from 1935-1960.
- "leading figure in establishing a national vital statistics system in the United States"

Record linkage is the task of assembling together all important pieces of information which refer to the same individual.

Record Linkage[*]

HALBERT L. DUNN, M.D., F.A.P.H.A.

*Chief, National Office of Vital Statistics, U. S. Public Health Service,
Federal Security Agency, Washington, D. C.*

EACH person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

The Book has many pages for some the various important records of a person's life.

The two most important pages in the Book of Life are the first one and the last one. Consequently, in the process of record linkage the uniting of the fact-of-death with the fact-of-birth has been given a special name, " death clearance."

Automatic Linkage of Vital Records*

Computers can be used to extract "follow-up"
statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

Proposed a probabilistic record linkage method and implemented it on the Datatron 205 computer.
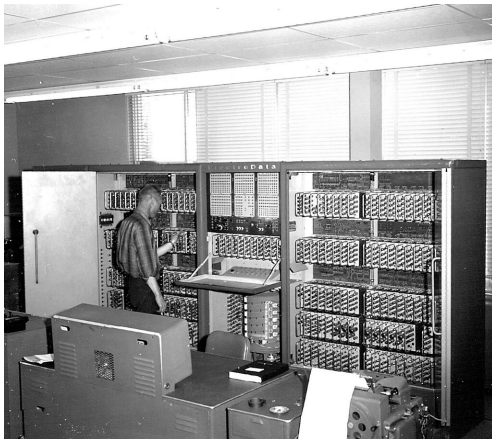
A THEORY FOR RECORD LINKAGE[*]

Ivan P. Fellegi and Alan B. Sunter

Dominion Bureau of Statistics

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

Fellegi and Sunter (1969), JASA

The authors formalized Newcombe et al. (1959) in a decision-theoretic framework.

One determines if two records are a match using a likelihood ratio test exceeding a threshold.

Fellegi and Sunter (1969), JASA

There are two methods proposed in this paper.

One method is completely unsupervised, and the other is semi-supervised.

The semi-supervised methods are used in practice for computational reasons.

Fellegi and Sunter (1969), JASA

Both the methods of Newcombe et al. (1959) and Fellegi and
Sunter (1969) have led to extensions that are utilized at statistical
agencies in order to update our "book keeping" regarding an
individual's book of life.

1. The methods must rely on training data.
2. There is sensitivity to tuning parameters such as the threshold
   used.
3. Both methods on their own do not scale to large data sets.
4. Transitive closures are not easily satisfied. See Sadinle and
   Fienberg (2013).

Modern Probabilistic Record Linkage

fastLink

The work of Enamorado et al. (2019) extends Fellegi and Sunter (1969) such that

1. the authors extend Lahiri and Larsen (2005) to incorporate auxiliary information such as population name frequency and migration rates into the merge procedure to conduct post-merge analyses

2. the authors are able to account for uncertainty of the merge process

3. the authors use parallelization and efficient data representations such that they can scale to millions of records

This work has been extended by Enamorado and **Steorts** (2020) to utilize fastLink as proposal for probabilistic blocking.

Bayesian FS Methods

Sadinle (2014) is a Bayesian Fellegi-Sunter method for de-duplication using a likelihood ratio similar in spirit to Fortini et al. (2001) and Fellegi and Sunter (1969).

The author considers a prior on the matching configuration matrix which imposes *transitive closures* — records are partitioned into groups which are thought to refer to the same entity.

This allows for uncertainty quantification via the posterior distribution.

The author provides corrections and a small labelled set of records for the UNTC data set as well a case study.

Bayesian FS Methods (Continued)

Sadinle (2017) extended this work to bipartite record linkage and derived Bayes estimates under a general class of loss functions, providing an alternative to the FS decision rule.

McVeigh et al. (2020) is an extension of Sadinle (2014, 2017) to include probabilistic blocking before the use of Bayesian FS.

The authors greatly improve the speed of the original approach of Sadinle and apply this methodology to a case study on voter registration and historical census records from California.

Semi-supervised and fully supervised methods

Semi-supervised methods use a relatively small amount of manually classified record pairs, known as labeled pairs, to improve upon unsupervised probabilistic record linkage.

Fellegi and Sunter (1969), Belin and Rubin (1995), Nigam et al. (2000), Larsen and Lahiri (2005), Chapelle et al. (2006), Christen et al. (2015), Kejriwal and Miranker (2015), Enamorado et al. (2019).

Fully supervised methods do not exploit information provided by unlabeled examples; instead they rely on larger amounts of labeled pairs.

1. Training data may come from approximate training sets (using unsupervised ER)
2. Training data may come by manual labelling of data
3. Training data may come from crowdsourcing extensive manual record linkage efforts.

Approximate training sets: Torvik et al. (2015)

Crowdsourcing: Sarawagi and Bhamidipaty (2002), Wang et al. (2012), Vesdapunt et al. (2014), Frisoli et al. (2019).

Manual efforts: Flemming et al. (2007), Christen (2007, 2008, 2014), Li et al. (2014), Ventura et al. (2014, 2015).

Clustering Based Methods

Why graphical Bayesian models?
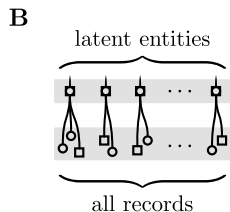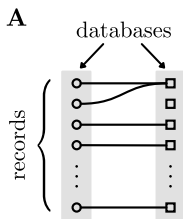
Tancredi and Liseo (2011), Steorts, Hall, Fienberg (2014, 2016), Steorts (2015), Zanella et al. (2016), Steorts, Tancredi, Liseo (2018), Marchant et al. (2020), Tancredi, Steorts, Liseo (2020), Betancourt et. al (2020)

distributed graphical entity resolution (`d-blink`)

Joint model for blocking and ER:

1. Scales to large databases using partially collapsed Gibbs sampling
2. Distributed inference/parallel inference is possible
3. Many computational speeds ups are proposed
4. Extensive studies on synthetic and real data sets are given
5. Open source software is provided in both Apache Spark and R.
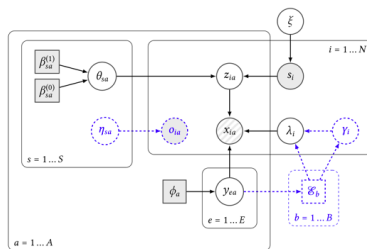
Marchant, Kaplan, Elazar, Rubinstein, **Steorts** (2020), JCGS, In Press. https://arxiv.org/abs/1909.06039

d-blink applied to synthetic and real data

Data set	Method	Pairwise measures			Cluster measures	
		Precision	Recall	F1-score	ARI	Err. # clust.
ABSEmployee	d-blink	0.9763	0.8530	**0.9105**	**0.9105**	+1.667%
	Fellegi-Sunter (10)	**0.9963**	0.8346	0.9083	—	—
	Fellegi-Sunter (100)	**0.9963**	0.8346	0.9083	—	—
	Near Matching	0.0378	**0.9930**	0.0728	—	—
	Exact Matching	0.9939	0.8346	0.9074	0.9074	+9.661%
NCVR	d-blink	0.9146	**0.9654**	**0.9393**	**0.9392**	−3.587%
	Fellegi-Sunter (10)	0.9868	0.7874	0.9083	—	—
	Fellegi-Sunter (100)	0.9868	0.7874	0.9083	—	—
	Near Matching	0.9899	0.7443	0.8497	—	—
	Exact Matching	**0.9925**	0.0017	0.0034	0.0034	+51.09%
NLTCS	d-blink	0.8319	0.9103	0.8693	0.8693	−22.09%
	Fellegi-Sunter (10)	**0.9094**	0.9087	**0.9090**	—	—
	Fellegi-Sunter (100)	**0.9094**	0.9087	**0.9090**	—	—
	Near Matching	0.0600	**0.9563**	0.1129	—	—
	Exact Matching	0.8995	0.9087	0.9040	**0.9040**	+2.026%
SHIW0810	d-blink	**0.2514**	0.5396	**0.3430**	0.3429	−37.65%
	Fellegi-Sunter (10)	0.0028	0.9050	0.0056	—	—
	Fellegi-Sunter (100)	0.0025	**0.9161**	0.0050	—	—
	Near Matching	0.0043	0.9111	0.0086	—	—
	Exact Matching	0.1263	0.7608	0.2166	0.2166	**−37.40%**
RLdata10000	d-blink	0.6334	**0.9970**	0.7747	**0.7747**	−10.97%
	Fellegi-Sunter (10)	0.9957	0.6174	0.7622	—	—
	Fellegi-Sunter (100)	0.9364	0.8734	0.9038	—	—
	Near Matching	0.9176	0.9690	**0.9426**	—	—
	Exact Matching	**1.0000**	0.0080	0.0159	0.0159	+11.02%

d-blink applied to the 2010 decennial census (Wyoming)

1. We consider the 2010 decennial census (Wyoming), which had a raw count of 563,626.
2. We merge the 2010 decennial census with administrative records from the Social Security Administration's Numerical Identification System (Numident).
3. We have partial ground truth via social security numbers.
4. Attributes: first and last name, gender, dob, and zip code.

Pairwise measures			Posterior population size	
Precision	Recall	F1-score	Mean	Std. error
0.97	0.84	0.90	616,000	5,000

Classical clustering

Many clustering tasks require models that assume cluster sizes grow linearly with the size of the data set.

Classical clustering

Many clustering tasks require models that assume cluster sizes grow linearly with the size of the data set.

Classic examples are the Dirichlet process (DP) and the Chinese Restaurant Process (CRP).

Classical clustering

Many clustering tasks require models that assume cluster sizes grow linearly with the size of the data set.

Classic examples are the Dirichlet process (DP) and the Chinese Restaurant Process (CRP).

More generally, we think of all infinite mixture models (Pitmor-Yor Process (PYP) and the Kingman Paintbox).
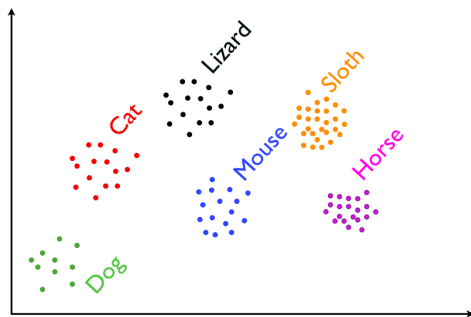
Flexible Models for Microclustering

How do we handle data where:

- We care about the exchangeability of the data points?
- But as the number of data points grows, the size of each cluster is negligible?

Microclustering

A sequence of random partitions $(C_N : N = 1, 2, \ldots)$ exhibits the *microclustering property* if M_N is $o_p(N)$, where M_N is the size of the largest cluster in C_N.

Microclustering

A sequence of random partitions $(C_N : N = 1, 2, \ldots)$ exhibits the *microclustering property* if M_N is $o_p(N)$, where M_N is the size of the largest cluster in C_N.

A clustering model exhibits the microclustering property if $(C_N : N = 1, 2, \ldots)$ implied by that model satisfies the above definition.

Zanella et al. (2016), Steorts et. al (2017), Johndrow et. al (2018), Betancourt, Zanella, Steorts (2020), Tancredi, Steorts, Liseo (2020).

Life After Entity Resolution:
Canonicalization and Downstream Tasks

Canonicalization, merging, or data fusion is the task of merging groups of records that have been classified as matches into one record that represents the true entity.

1. The earliest proposals of canonicalization were deterministic, rule-based methods, which were application specific and fast to implement (Cohen and Sagiv, 2005).

2. The existing literature assumes training is available in order to select the canonical record, and authors have proposed optimization and semi-supervised methods to find the most representative record.

3. Motivated by the NCSBE voters data set, Kaplan et al. (2020) provide a unique identifier for voter registration in a principled and reproducible manner.

4. For a full review of data fusion techniques, we refer to Bleiholder and Naumann, 2009.

Christen (2012), Culotta et al (2007), Bleiholder and Naumann (2009).

Turning to joint or single-stage modeling approaches to entity resolution and the downstream task, these have been limited to linking two databases and do not easily generalize beyond this framework.

Recent applications to human rights data have been done by Sadinle (2018) and Tancredi, Steorts, and Liseo (2020) involving population sized estimation (multiple systems estimation).

Single-stage, joint models, and canonicalization tasks cover a vast body of literature. For a more thorough review, be on the look out for a review paper by Binette and Steorts (2020) on entity resolution.

Open Research Problems and Discussion Questions

1. How should one do fair comparisons and evaluations?
2. How should one create training data sets?
3. How should one find publicly available benchmark data sets and which ones should be used?
4. Should we be aware that entity resolution can be used for massive data collection efforts and evasion of privacy? What can we do as a field regarding this?
5. What privacy guarantees are offered for record linkage?
6. What ethical considerations do I have when working with private data that is sensitive, such as the data that HRDAG provides?
7. What resources are available to me if I want to know more about the field of record linkage?
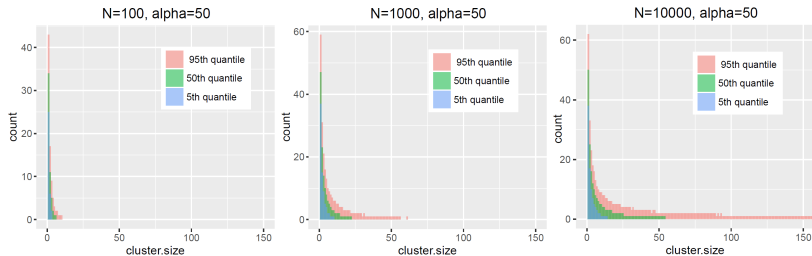
Questions?

beka@stat.duke.edu

Webpage: `resteorts.github.io`

Software: `https://github.com/orgs/cleanzr/`

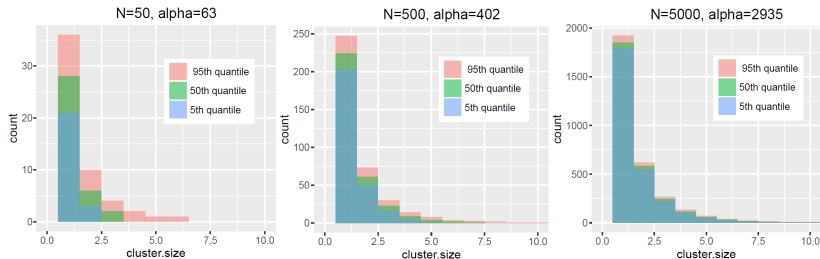Paper: `https://arxiv.org/abs/2008.04443`

CRP Behavoir for N increasing and α fixed

As N increases the size of the clusters increases as $O(N) \implies$ Not appropriate for microclustering.

CRP Behavoir for N and α jointly increasing

Can obtain microclustering property by increasing α with N but the resulting model becomes less flexible.

The model collapses onto a one parameter family distribution for the cluster sizes.

This behavior motivates models that naturally incorporate the microclustering property.