# Many Eggs, More Baskets: New Insights from New Models

## Thesis Proposal

Clément Canonne

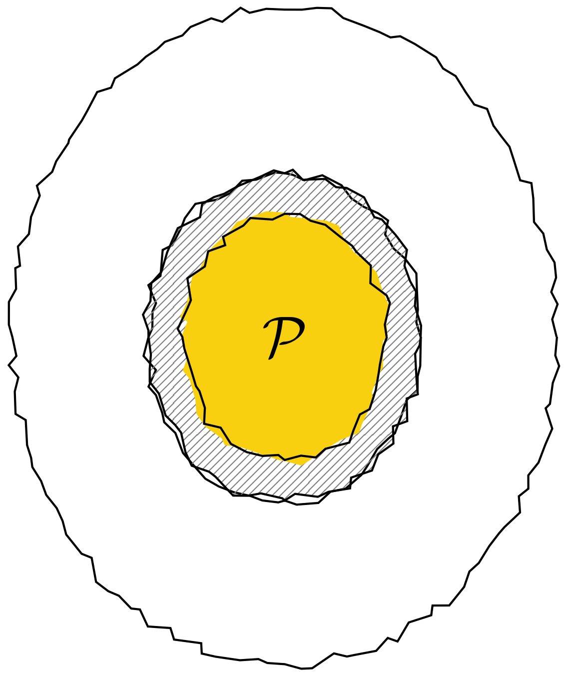
Columbia University – 2016

# Introduction

# Introduction

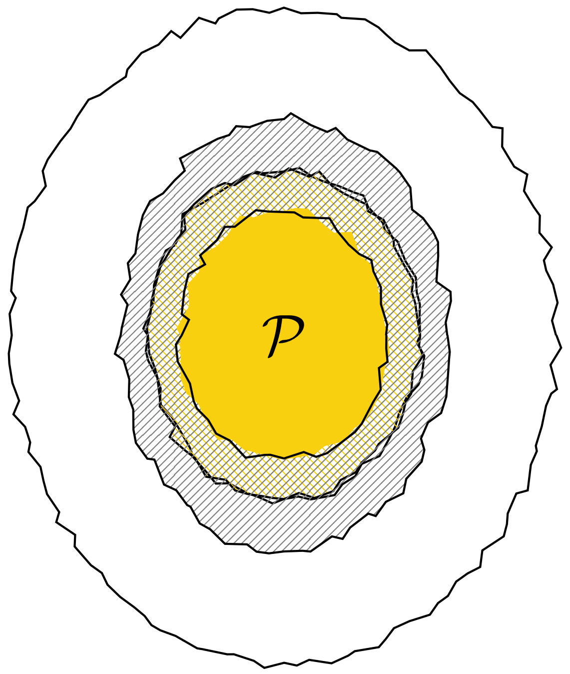


**Property testing:** what can we say about an object while barely looking at it?



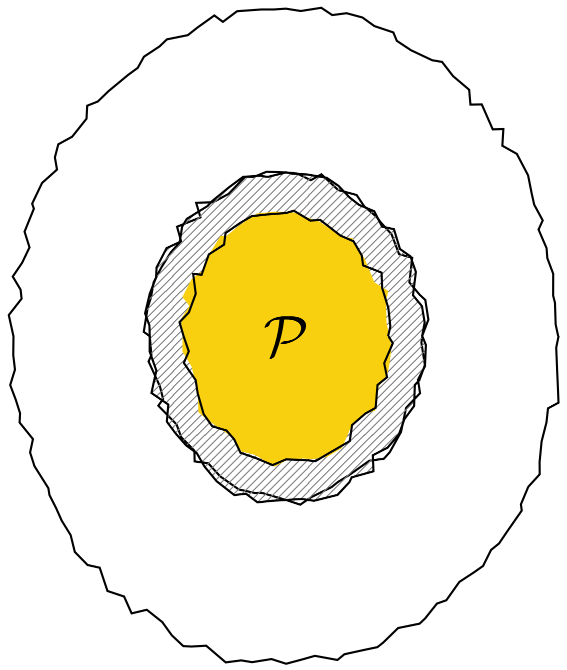

"Is it in the yolk?"

**Tolerant testing:** robust version of property testing




**(Typically harder.)**

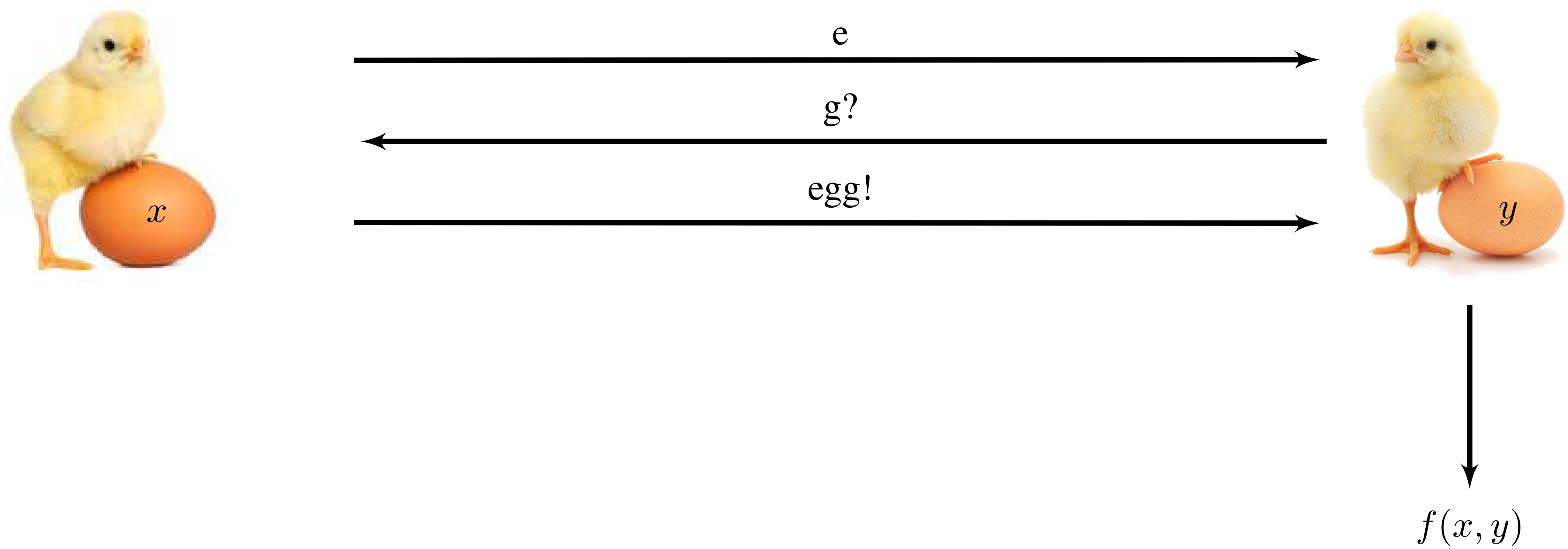
**Distribution testing:** property testing for probability distributions




**Different metric, objects, and type of access.**

# Introduction

**Communication complexity:**



# Outline of the talk

# In and Beyond Distribution Testing

# The standard setting

$\Delta(\Omega)$: all distributions over (finite) domain $\Omega$ of size $n$. **Property:** subset $\mathcal{P} \subseteq \Delta(\Omega)$. **Tester:** randomized algorithm (knows $n$, $\mathcal{P}$).

Given independent samples from a distribution $D \in \Delta(\Omega)$, and parameter $\varepsilon \in (0,1)$, output accept or reject:

- If $D \in \mathcal{P}$, accept with probability at least 2/3; *(in the yolk)*
- If $\ell_1(D, \mathcal{P}) > \varepsilon$, reject with probability at least 2/3; *(definitely white)*
- otherwise, whatever (make an omelet).

**Goal:** take $o(n)$ samples, ideally $O_\varepsilon(1)$.
(time efficiency is secondary, yet not frowned upon.)

[BFF+01, BKR04, BFR+10, GGR98]

# The challenges



**Unified frameworks:** how to get past the *ad hoc*, property-specific results (upper and lower bounds) to get *generic* approaches?

**Strong lower bounds:** how to get around the hardness results in the standard sampling model (e.g. [VV10a])?

**Strong assumptions:** how to get rid of (some) of the assumptions – can we deal with *limited independence*?

# Several chickens with one stone

# Several chickens with one stone



Many individual results on specific properties:

- Uniformity
- Identity
- Equivalence
- Independence
- Monotonicity
- Poisson Binomial Distributions
- and more…

…but almost none on general frameworks.

# Several chickens with one stone

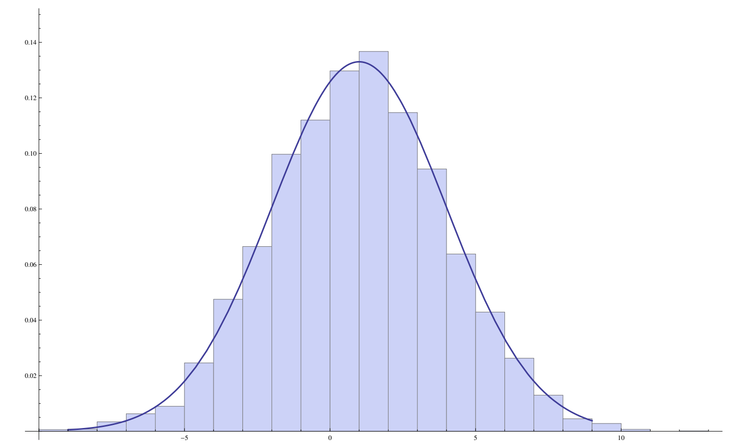
How to get past the *ad hoc*, property-specific results (upper and lower bounds) to get *generic* approaches?

**1. Abstract structural properties of the properties.**

# Several chickens with one stone

A unified approach to things:

- define a structural criterion (parameterized by some quantity $L$) of classes of distributions



- obtain a single testing algorithm $\mathcal{T}$ that takes $L$ as input

$$D \rightsquigarrow \mathcal{T}(L) \rightsquigarrow \text{accept/reject}$$

- Prove existential result for your favorite class $\mathcal{C}$:

$$\mathcal{C} \rightsquigarrow L(\mathcal{C}, \varepsilon)$$

- Use $\mathcal{T}$ to test $\mathcal{C}$

$$D \rightsquigarrow \mathcal{T}(L(\mathcal{C}, \varepsilon)) \rightsquigarrow \text{accept/reject}$$

# Several chickens with another stone

How to get past the *ad hoc*, property-specific results (upper and lower bounds) to get *generic* approaches?

**2. Do the (supposedly) impossible.**

# Several chickens with another stone

Other generic frameworks:

- Upper bounds by learning-and-testing [ADK15, Can16]

$$\chi^2 \leq \varepsilon^2 \text{ vs. } \mathrm{d}_{\mathrm{TV}} > \varepsilon$$

- Upper bounds *via* $\ell_2$ testing and randomized mapping [DK16]

$$D \in \Delta([n]) \rightsquigarrow F(D) \in \Delta([N]) \rightsquigarrow \ell_2\text{-testing}$$

- Lower bounds *via* blackbox reductions [CDGR15]

$$\mathcal{C}^{\text{Hard}} \subseteq \mathcal{C} \rightsquigarrow \text{testing}(\mathcal{C}^{\text{Hard}}) \preceq \text{testing}(\mathcal{C})$$

- Lower bounds *via* information theory [DK16]

# Several chickens with a big rock

How to get past the *ad hoc*, property-specific results (upper and lower bounds) to get *generic* approaches?

**3. Ask Alice and Bob.**

# Communication Compleggsity

# Several chickens with a big rock

Approach *à la* [BBM11]: reduction from communication complexity:

- select the right communication setting:
$$A \to B, \quad A \leftrightarrow B, \quad A \to R \leftarrow B \dots$$
- choose a hard enough communication problem:
$$\textsc{Disjointness}, \textsc{Gap-Hamming}, \textsc{Equality}^*, \text{something new} \dots$$
- create distance from the CC inputs:
$$(a,b) \in \mathcal{Y} \rightsquigarrow D_{a,b} \in \mathcal{P} \qquad (a,b) \in \mathcal{N} \rightsquigarrow \ell_1(D_{a,b}, \mathcal{P}) > \varepsilon$$
- simulate access from the CC inputs
$$A \to B \rightsquigarrow s \sim D_{a,b}$$

# Strengthening the oracle

# Strengthening the oracle

How to get around the hardness results in the sampling model?

**Question the model.**

# Strengthening the oracle

Changing the model of access to $D$:

- with evaluation queries to the pmf: [RS09] ("property-testing"-style)

$$x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the pmf: [BDKR05, GMV06, CR14]

$$? \rightsquigarrow x \sim D \quad \text{and} \quad x \in \Omega \rightsquigarrow D(x)$$

- with sampling and evaluation queries to the cdf: [BKR04, CR14, Can15]

$$? \rightsquigarrow j \sim D \quad \text{and} \quad j \in [n] \rightsquigarrow \sum_{i=1}^{j} D(i)$$

- with conditional sampling: [CFGM13, CRS15, ADK15, Can15]

$$S \subseteq \Omega \rightsquigarrow x \sim D_S$$

# Results: the Sunny Side (Up)

**Informally:** across the models and flavors, exponential sample complexity improvements – sometimes even from $n^{\Omega(1)}$ to constant. Some hardness remains, still – and most importantly, *all rules of thumbs are down.*

**Conditional sampling:** identity and closeness testing are no longer related ($O_\varepsilon(1)$ vs. $(\log\log n)^{\Omega(1)}$). Tolerant uniformity testing and entropy estimation are, similarly, worlds apart.

**Testing with queries:** Testing uniformity, identity and closeness becomes easy: the challenge now seems to lie in tolerant testing, or in testing against classes.

**Challenges:** Understanding how these new models relate, and develop generic tools to analyze them.

# Weakening the assumptions

# Weakening the assumptions

How to get rid of (some) of the assumptions – can we deal with limited independence?

**Semi-adversarial setting, capturing real-life situations: memory pages, hard drive, clustered data…**

# Weakening the assumptions

Work in the external memory model of [AIOR09]:

- multiset $S \subseteq [n]$ of $m$ datapoints, clustered (arbitrarily) in blocks of size $B$

$$|S_1| = |S_2| = \cdots = |S_{m/B}| = B$$

- random access to the blocks, reading a full block has unit cost

$$i \rightsquigarrow S_i$$

- want to test properties of the dataset: of $D \in \Delta([n])$ induced by $S$

$$D(i) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{\{s_j = i\}}$$

- take advantage of this egg in our beer: optimal in $m, B, \varepsilon, n$?

# Other and Future work

# Other and Future work

Other “neglected” or novel settings: what fails to be addressed or captured - and ought to be?

**Imperfect communication, Sampling correction, Robust function testing**

# Proposed Timeline

| Timeline | Work | Progress |
| --- | --- | --- |
| Sep. 2012–May 2016 | Unified testing, Conditional sampling (COND), Extended access model, Sampling correctors, Communication with Imperfect randomness (ISR) | completed |
| Spring 2016 | Submit full version of ISR to IEEE-IT | in review |
| Spring 2016 | Submit full version of (second) COND to ToC | in review |
| Spring 2016 | Submit full version of Sampling Correctors to SICOMP | in progress |
| Jan. 2016–July 2016 | Limited Independence | in progress |
| Oct. 2015–Aug. 2016 | Lower bounds via CC | in progress |
| Dec. 2015–Oct. 2016 | Tolerant Junta testing | in progress |
| Fall 2016 | Followup results on conditional sampling? | |
| Dec. 2016–March 2017 | Thesis writing | |
| Spring 2017 | **Thesis defense** | |
| *Spring 2017–Forever* | *Stay here?* | *(wishful)* |

# The ~~End~~

# Bibliography (1)


[ACK15] J. Acharya, C. L. Canonne, and G. Kamath. A chasm between identity and equivalence testing with conditional queries. In *RANDOM*, 2015.

[AD14] J. Acharya and C. Daskalakis. Testing Poisson Binomial Distributions. In *SODA*, 2014.

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *NIPS*, 2015.

[AIOR09] Alexandr Andoni, Piotr Indyk, Krzysztof Onak, and Ronitt Rubinfeld. External sampling. In *ICALP*, 2009.

[BBM11] Eric Blais, Joshua Brody, and Kevin Matulef. Property testing lower bounds via communication complexity. In *CCC*, 2011.

[BDKR05] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SICOMP*, 35(1):132–150, 2005.

[BFF$^+$01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *FOCS*, 2001.

[BFR$^+$10] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. (abs/1009.5397), 2010.

[Bir87] L. Birgé. On the risk of histograms for estimating decreasing densities. *The Annals of Statistics*, 15(3), 1987.

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *STOC*, 2004.

[Can15] Clément L. Canonne. Big Data on the rise? Testing monotonicity of distributions. In *ICALP*, 2015.

[Can16] C. L. Canonne. Are Few Bins Enough: testing Histogram Distributions In *PODS*, 2016.

[CDGR15] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing Shape Restrictions, 2015. STACS'16.

[CDVV14] S-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, 2014.

[CFGM13] S. Chakraborty, E. Fischer, Y. Goldhirsh, and A. Matsliah. On the power of conditional samples in distribution testing. In *ITCS*, 2013.

[CR14] C. L. Canonne and R. Rubinfeld. Testing probability distributions underlying aggregated data. In *ICALP*, 2014.

[CRS15] C. L. Canonne, D. Ron, and R. A. Servedio. Testing probability distributions using conditional samples. *SICOMP*, 44(3):540–616, 2015.

[DDS12] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning $k$-modal distributions via testing. In *SODA*, 2012.

# Bibliography (2)



[DDS+13] C. Daskalakis, I. Diakonikolas, R. A. Servedio, G. Valiant, and P. Valiant. Testing $k$-modal distributions: Optimal algorithms via reductions. In *SODA*, 2013.

[DKN15] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In *SODA*, 2015.

[DK16] I. Diakonikolas and D. M. Kane. A New Approach for Testing Properties of Discrete Distributions. Manuscript, 2016.

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *JACM*, 45(4):653–750, July 1998.

[GMV06] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA*, 2006.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, ECCC, 2000.

[LRR13] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. *Theory Comput.*, 9:295–347, 2013.

[LRR14] R. Levi, D. Ron, and R. Rubinfeld. Testing similar means. *SIDMA*, 28(4):1699–1724, 2014.

[Pan04] L. Paninski. Estimating entropy on $m$ bins given fewer than $m$ samples. *IEEE-IT*, 50(9), 2004.

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE-IT*, 54(10), 2008.

[PRR06] M. Parnas, D. Ron, and R. Rubinfeld. Tolerant property testing and distance approximation. *JCSS*, 72(6):1012–1042, 2006.

[RRSS09] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distributions support size and the distinct elements problem. *SICOMP*, 39(3):813–842, 2009.

[RS09] R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. *RSA*, 34(1):24–44, January 2009.

[Val11] P. Valiant. Testing symmetric properties of distributions. *SICOMP*, 40(6):1927–1968, 2011.

[VV10a] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. *ECCC*, 17:179, 2010.

[VV10b] G. Valiant and P. Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *ECCC*, 17:180, 2010.

[VV11] G. Valiant and P. Valiant. The power of linear estimators. In *FOCS*, 2011. See also [VV10a] and [VV10b].

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.