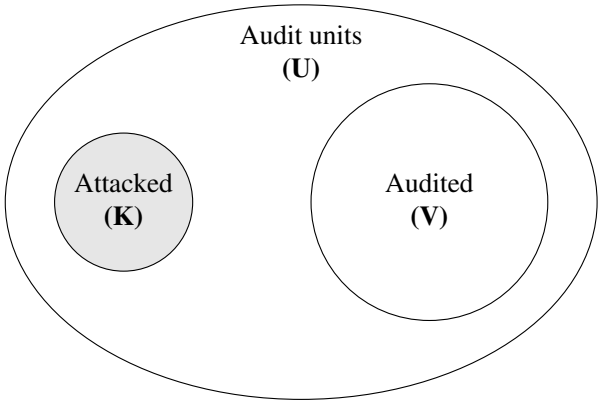
On the Security of Election Audits with Low Entropy Randomness

Eric Rescorla
ekr@rtfm.com

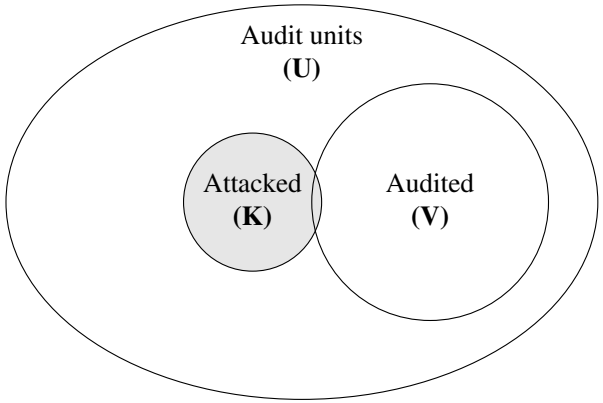
Overview

- Secure auditing requires random sampling
  - The units to be audited must be verifiably unpredictable
  - Simple physical methods (dice, coins, etc.) are expensive
- "Stretching" approaches
  - Randomness tables [CWD06]
  - Cryptographic pseudorandom number generators (CSPRNGs) [CHF08]
- These techniques must be seeded with verifiably random values
- Small (but natural) seeds give the attacker an advantage

Formalizing the Problem: The Auditing Game



$\mathbf{V} \cap \mathbf{K} = \emptyset$: Attacker wins



$\mathbf{V} \cap \mathbf{K} \neq \emptyset$: Attacker loses

- Two players: Attacker and Auditor
- $\mathbf{U}$ audit units $(U_0, U_1, ...U_{N-1})$
- Attacker selects $\mathbf{K} \subset \mathbf{U}$ to attack $(|\mathbf{K}| = k)$
  - Selection is made before preliminary results are posted
- Auditor selects $\mathbf{V} \subset \mathbf{U}$ to audit $(|\mathbf{V}| = v)$

Auditing Game Strategy

- If the auditor's selections are random and i.i.d then:

$$\Pr(\text{detection}) = 1 - \prod_{i=0}^{v-1} \frac{(N-i-k)}{N-i}$$

- No matter how the attacker chooses $\mathbf{K}$
- This is the auditor's optimal strategy
- What about intermediate cases?
  - Attacker has incomplete information about $\mathbf{V}$

Example: A Million Random Digits [RAN02]

TABLE OF RANDOM DIGITS 1

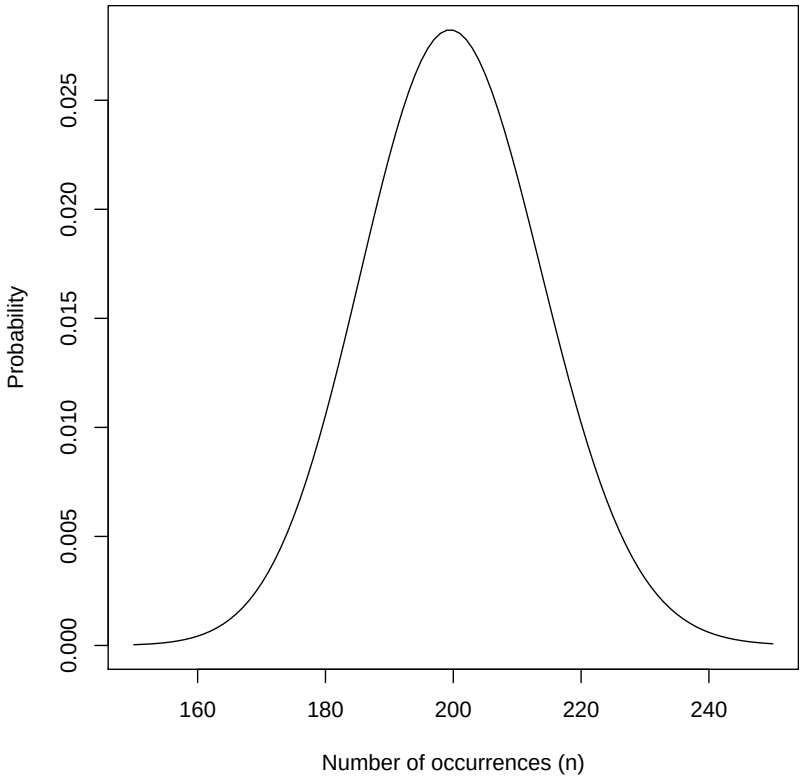
00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790

- Pick a random starting group and read forward
  - This process has $log_2(\#entries)$ bits of entropy

Random Number Tables Bias and Attacker Advantage

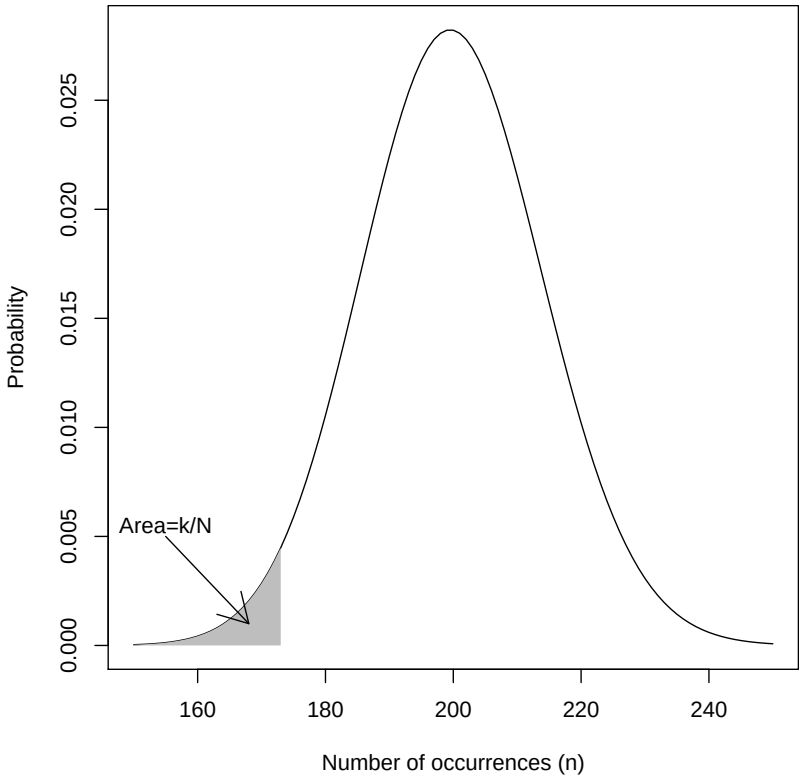
- Random number tables aren't the same as random numbers
  - The attacker knows the table
  - But not the starting point
- Two effects give the attacker an advantage
  - Natural variation in the occurrences of each value
  - Clustering of values

Natural Variation



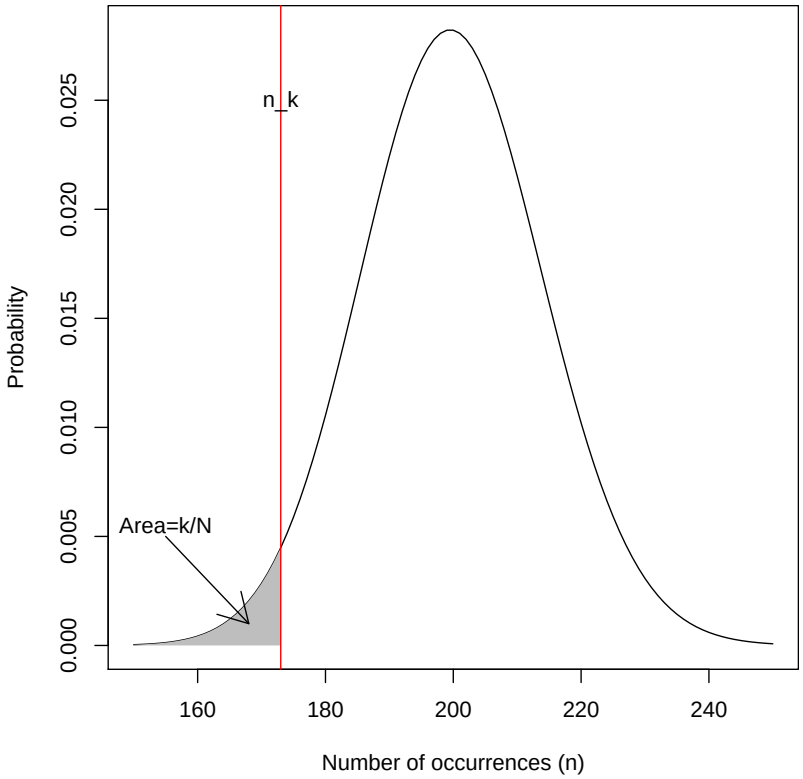
- Binomially distributed counts
- Expected value $= T/N$

Natural Variation



- Binomially distributed counts
- Expected value $= T/N$
- Attacker selects k least frequent units

Natural Variation



- Binomially distributed counts
- Expected value $= T/N$
- Attacker selects k least frequent units
- The kth least frequent unit appears n_k times

$$n_k = \min\left\{n : \mathrm{cdf}(n) \geq \frac{k}{N}\right\}$$

Auditing with Natural Variation

- Total entries in table corresponding to k least frequent units[†]:

$$T_{bad} = N \sum_{n=0}^{n_k} n\varphi(n)$$

- This is just a standard sampling problem
  - Each "good" sample removes approximately F entries:

$$F = \frac{T - T_{bad}}{N - k}$$

- Probability of detection of least frequent k units:

$$\Pr(\text{detection}) = 1 - \prod_{i=0}^{v-1} \frac{T - iF - T_{bad}}{T - iF}$$

[†]Semi-accurate approximation; see paper.

Clustering Effects

- We're not really sampling the table randomly
  - We read entries in sequence
  - The order of the entries matters

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
6	7	8	9	2	3	4	5	6	7
8	9	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	2	3
4	5	6	7	8	9	2	3	4	5
6	7	8	9	2	3	4	5	6	7
8	9	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	2	3
4	5	6	7	8	9	2	3	4	5

A table constructed to minimize detection

0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9
0	2	3	4	5	1	6	7	8	9

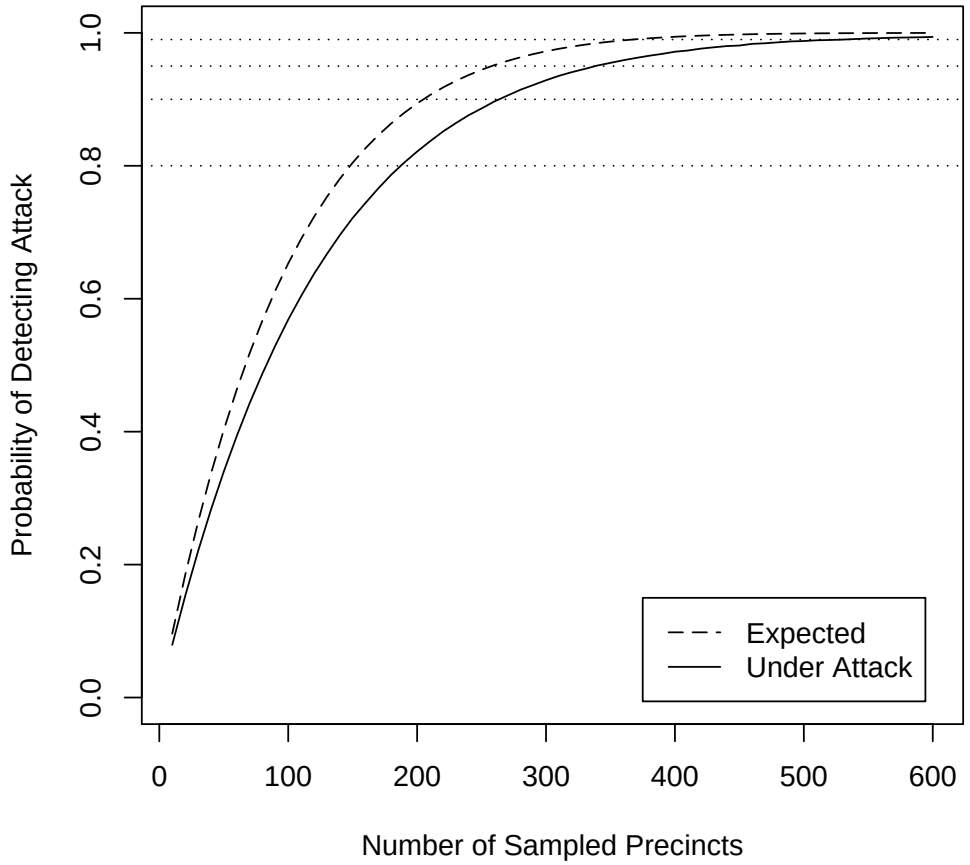
A table constructed to maximize detection

Simulation Studies

- No good analytic model for clustering effect
  - Though some potential avenues
- Easiest to study via simulation
  - Generate a random table (using CSPRNG)
  - Generate an attack set of size k
  - Determine which offsets will sample at least one element of $\mathbf{K}$
- Two kinds of attack sets
  - Random (should have expected statistics)
  - Randomly selected from least frequent $2k$ units[†]
- Results averaged over multiple tables (5–25)

[†] This is heuristic. We don't have a good algorithm here either.

Example



200,000 entries, 1000 precincts, 10 attacked

The Attacker's View: Modest Advantage

- Still very likely to be detected
  - In the above example: about 4x more chance of success at 99%
  - Biggest gap around 80% nominal detection rate (71.4% actual)
- Probably not enough to make or break an attack
  - But worth doing if you're going to attack anyway

The Auditor's View: Higher Work Factor

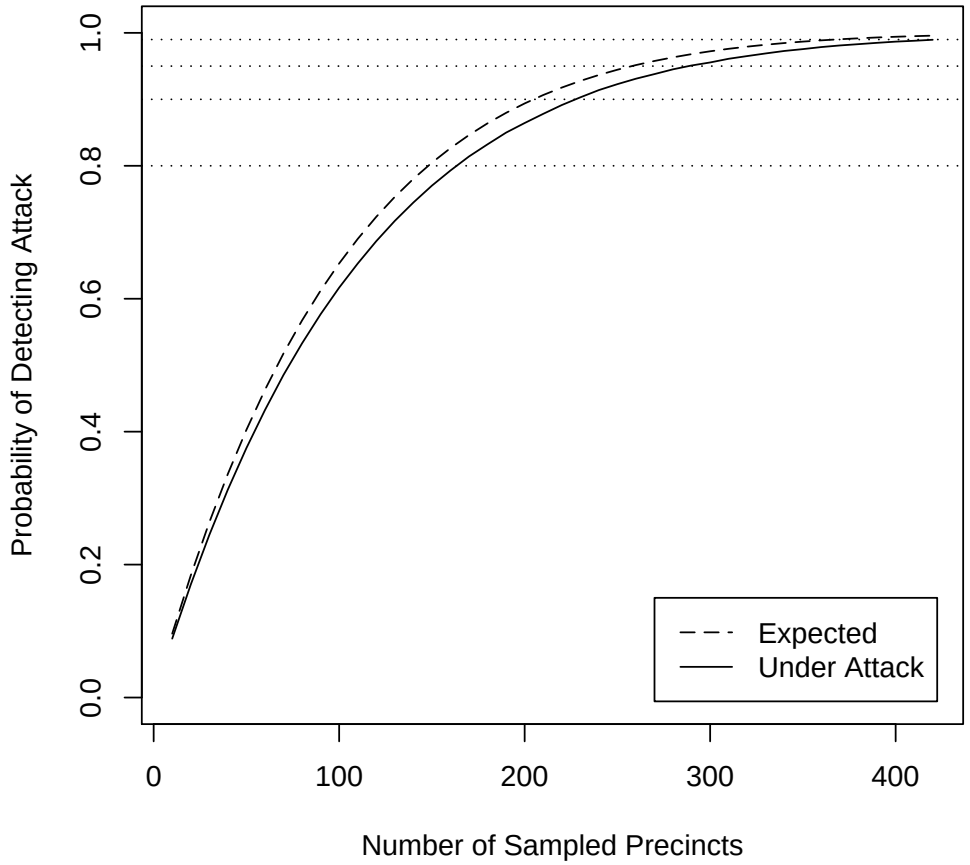
Detection Probability	Units to Audit (projected)	Units to Audit (under attack)†	Difference (percent)
80%	148	190	28
90%	205	270	32
95%	258	340	32
99%	368	540	47

Required audit levels: 200,000 entries, 1000 precincts, 10 attacked precincts

General Trends

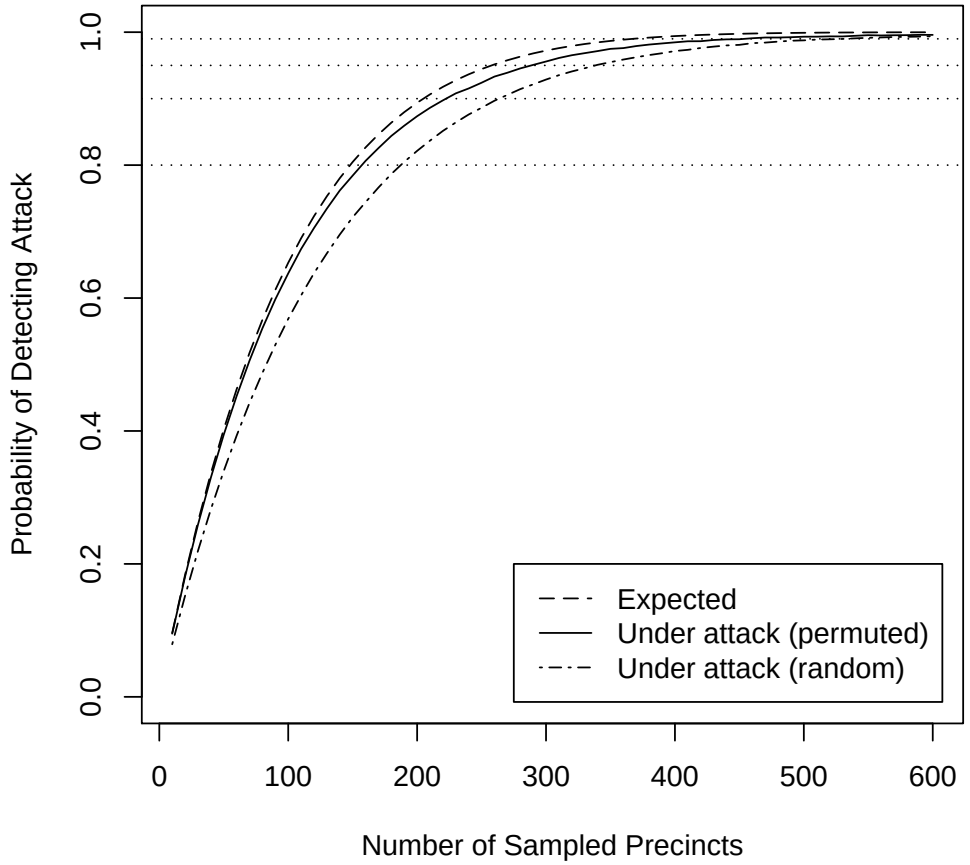
- More entries per unit decrease attacker advantage
  - Larger tables
  - Fewer units
- Higher attack rates decrease attacker advantage
  - Need to select increasingly probable values

A Big Table



1,000,000 entries, 1000 precincts, 10 attacked precincts

Permuted Tables



200,000 entries, 1000 precincts, 10 attacked precincts

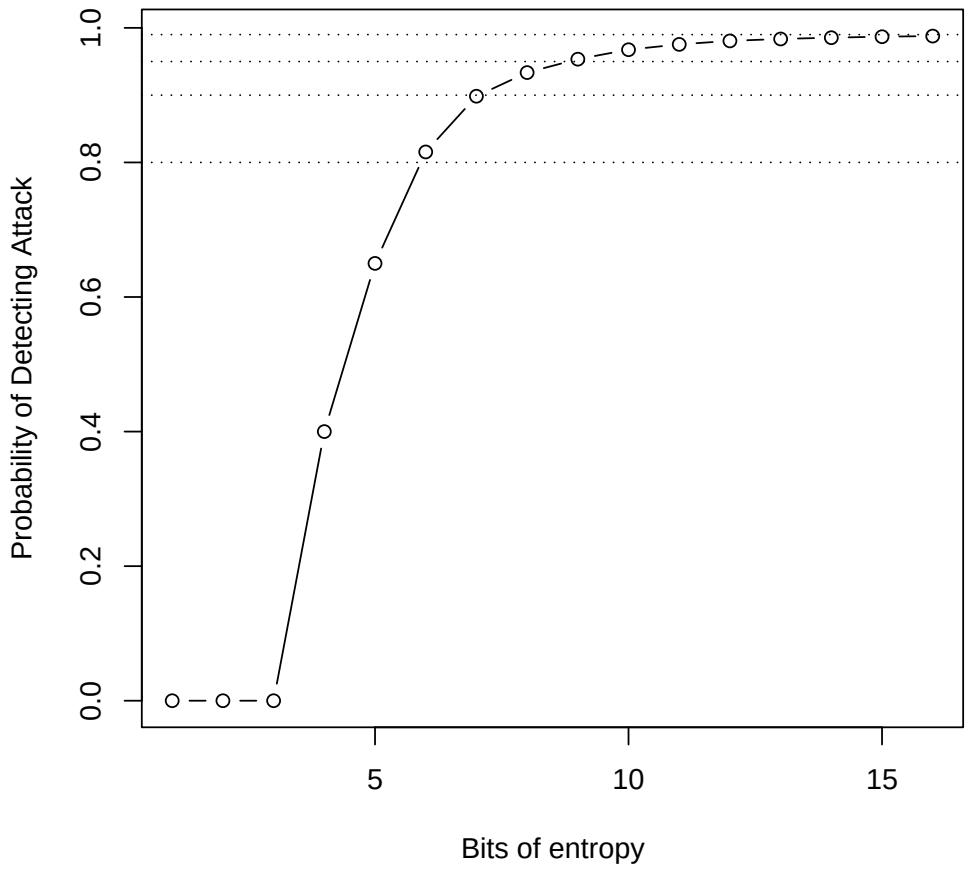
Potential Improvements

- New tables
  - Bigger (10^7 entries?)
  - Permuted rather than random
  - Generated using a PRNG?
- Existing tables
  - Individual addressing
  - Random offsets
  - Multiple starting points
  - All of these need analysis

What about CSPRNGs?

- CSPRNGs have big state spaces no matter what the seed size
  - Stronger than tables for the same seed entropy
  - Intuition: sequences don't overlap
- Cryptographic applications require very large seeds
  - Not necessary here
  - Need unpredictability, not unsearchability

Security of PRNGs by Seed Size (nominal 99% level)



Probability of detection for PRNGs: 1000 precincts, 10 attacked

Summary

- Secure auditing requires verifiably unpredictable random values
- Generating them directly seems expensive
- Natural stretching approaches may not deliver their expected security
- Not clear if randomness tables can be used safely
- PRNGs appear safe with modest-sized seeds

References

[CHF08] Joseph A. Calandrino, J. Alex Halderman, and Edward W. Felten. In Defense of Pseudorandom Sample Selection. In Proceedings of the 2008 Electronic Voting Technology Workshop, 2008. http://www.usenix.org/events/evt08/tech/full_papers/calandrino/calandrino.pdf.

[CWD06] Arel Cordero, David Wagner, and David Dill. The role of dice in election audits—extended abstract. IAVoSS Workshop on Trustworthy Elections 2006 (WOTE 2006), June 2006. http://www.cs.berkeley.edu/~daw/papers/dice-wote06.pdf.

[RAN02] RAND Corporation. A Million Random Digits with 100,000 Normal Deviates. American Book Publishers, 2002.