Public Voting Records: A Record, or an Attack Surface?

A formal methodology for auditing voter-file disclosure regimes against linkage attacks

Noah M. Kenney

Founder & Principal Consultant, Digital 520

Abstract. *Every U.S. state publishes voter registration data, but states differ on which fields they release. Texas withholds date of birth, race, party affiliation, and phone number; North Carolina publishes all four. Conventional wisdom holds that the more conservative regime substantially protects voter privacy. We test that wisdom empirically. We formalize a four-step methodology, a linkage map of every external dataset that joins to the file, a linkage ladder that reports the share of voters identifiable under each plausible attacker input, a chained-information-gain table, and a harm taxonomy with seven elaborated case studies, and prove three theoretical properties of the linkage ladder (monotonicity in the attacker's knowledge set, intersective composition under disjoint quasi-identifiers, and a collision-probability lower bound on uniqueness) before applying the framework to Travis County, TX (n = 879,827) and Robeson County, NC (n = 63,435 active). Three empirical findings stand out. First, both files are more than 85% re-identifiable under trivial attacker inputs: name and ZIP code together identify 95.81% of Travis voters uniquely (Wilson 95% CI [95.77, 95.85]) and 87.79% of Robeson voters; neither regime substantially protects against this attack. Second, Texas's DOB redaction is largely offset by an under-recognized substitution effect: registration date, published at single-day resolution, plays the same identifying role that DOB would. Generalizing it to year resolution would reduce the relevant uniqueness rate by a factor of 558 with no loss of utility. Third, robustness analyses confirm the headline statistics are stable to confidence-interval treatment, geographic skew across 58 ZIPs, four name-normalization variants, and three error-injection rates; counter-intuitively, data errors slightly increase rather than decrease the attacker's success rate. We further demonstrate the attack empirically against the FEC contribution database, achieving a 52.49% unique-voter match rate without name normalization. We argue that field-level disclosure is empirically a secondary determinant of contemporary voter-file privacy outcomes and that access-control mechanisms (rate limiting, requester verification, audit logging, resale prohibitions) are the dominant lever; we propose a corresponding shift in policy emphasis.*

Index Terms. *voter privacy, re-identification, k-anonymity, linkage attack, quasi-identifier, public records, disclosure policy, threat modeling, doxxing, data brokers, election security, adversary model, robustness analysis.*

I. Introduction

Every state in the United States requires public release of voter registration data in some form. The conventional understanding of voter-file privacy frames the question as which fields to publish: name and address are universal; date of birth, race, party affiliation, and phone number are subject to state-by-state variation. Texas, taking the conservative position, withholds the latter four. North Carolina publishes them all. The two states therefore anchor opposite ends of the U.S. disclosure spectrum, and we use them as our two empirical anchors throughout this paper.

The contribution of this paper is to formalize the analysis. Prior work in re-identification, beginning with Sweeney [1], reports point estimates of population uniqueness under fixed quasi-identifier triples (most famously ZIP + DOB + sex → 87% unique). The contribution of those estimates was foundational, but the policy debate that grew up around them has substantially mistaken the structure of the underlying problem: redaction of one quasi-identifier is wasted effort if another field of comparable entropy remains; the field-level disclosure question is not the binding constraint on contemporary voter-file privacy; and the absence of a unified adversary model has made it difficult to compare jurisdictions or evaluate proposed reforms on a common axis.

We make seven contributions. (1) We formalize a four-step methodology consisting of a linkage map, a linkage-ladder function defined over quasi-identifier subsets, a chained marginal-information-gain table, and a harm taxonomy with case studies. (2) We prove three theoretical properties of the linkage ladder, monotonicity under set inclusion, intersective composition under disjoint quasi-identifiers, and a collision-probability lower bound on uniqueness, that hold for any voter file or any public PII dataset to which the methodology is applied. (3) We introduce a canonical adversary model parameterized by knowledge set, budget, capability set, and harm intent, and recast each threat-model case study in this notation. (4) We apply the methodology to two voter files anchoring opposite ends of the U.S. disclosure spectrum, with all numeric claims reproducible from a single ~25-second Python script. (5) We isolate a quasi-identifier substitution effect (the EDRDAT-as-DOB-proxy result) which explains why field-level redaction in Texas does not produce the privacy benefits commonly attributed to it. (6) We report robustness analyses including Wilson confidence intervals, per-ZIP

variance, four name-normalization variants, and three character-level perturbation rates. (7) We run a real-data linkage to the FEC contribution database, producing an empirical lower bound on the theoretical re-identification rate.

The principal empirical finding we build toward is that even a voter file at the conservative end of the U.S. disclosure spectrum (Texas) admits k=1 re-identification by trivial attacker knowledge sets at rates exceeding 75% (name alone) and 95% (name + ZIP), and that the file at the permissive end (North Carolina) yields rates that are quantitatively similar. The substantive policy implication is that field-level redaction is an empirically secondary determinant of contemporary voter-file privacy outcomes, not the dominant lever the contemporary debate has treated it as. We argue this analytically: the ladder is monotone, the empirical curves cross at a point well above the relevant action threshold, and the marginal value of further field-level redaction is bounded by the linkability already supplied by name and geography.

II. Research Questions

This paper addresses four research questions, each operationalized by a specific empirical procedure within the formal framework introduced in §IV.

RQ1: Given a published voter file, what other datasets does it allow an adversary to pull in? Operationalization: structured catalog of candidate external datasets along five dimensions (dataset class, join keys F, new attributes revealed, access cost, prior-art harms). Output: the linkage map (§IX).

RQ2: What are the risks of linking those datasets to publicly available information? Operationalization: each linkage is mapped to a structured harm taxonomy with six modalities, and seven case studies elaborate one or more modalities each (§X).

RQ3: What percentage of voters can be linked? Operationalization: for each quasi-identifier subset F that an attacker plausibly possesses, compute the linkage-ladder values $L\_D(F) = (\Pr[k\_F = 1], \Pr[k\_F < 5])$. Output: linkage ladders for both files (§VII-B, §VII-C). One entry of the ladder is empirically validated by a real join to the FEC contributions database (§VII-F).

RQ4: What is the worst harm that can come from this? Operationalization: case studies in §X, each grounded in a vulnerable-class size computed from the data, an adversary tuple specified in the formal model, a budget estimate, and a verified prior-art incident.

III. Background and Related Work

A. Foundational re-identification work

The foundational result is Sweeney's 2000 finding that 87% of the U.S. population can be uniquely identified by ZIP, date of birth, and sex [1]. The result was demonstrated against the Massachusetts Group Insurance Commission discharge file by joining to the publicly available Cambridge, MA voter list, ultimately re-identifying then-Governor William Weld's medical record. Sweeney subsequently formalized k-anonymity [2] and demonstrated re-identification of 84 to 97% of Personal Genome Project profiles [3]. Sweeney, Yoo, and colleagues showed in 2017 that voter-registration tampering attacks succeed against websites in 35 states + DC using publicly available identifiers [4]. The k-anonymity model has been extended by l-diversity, t-closeness, and differential privacy in subsequent work; we draw on the original k-anonymity formulation because it most directly captures the linkage-attack semantics relevant to voter-file disclosure.

B. Data-broker and people-search literature

A parallel literature documents harms from commercial data brokers and people-search services, which ingest public records (including voter files) and resell them. The Amy Boyer murder of 1999 [6] is canonical: Boyer's stalker paid Docusearch $45 for an SSN and $109 for a work address, then killed her. Remsburg v. Docusearch (NH 2003) established broker duty of care but did not modify the disclosure regimes that fed the broker. The Lawfare data-broker series [5] catalogues the "publicly available information" carve-outs that exempt brokers from many state privacy laws so long as the data originated in a public record.

C. Recent harms

Post-Dobbs (2022), threats against reproductive-care workers rose 20% and stalking incidents rose 229%, with doxxing as a primary modality [8]. The U.S. Government Accountability Office reported in October 2025 [9] that publicly accessible digital data, voter files among other sources, aggregates into "digital profiles" exposing service members. The Burkman and Wohl 2020 robocall scheme [10] targeted approximately 85,000 voters across Michigan, Illinois, New York, Pennsylvania, and Ohio; both defendants pleaded no contest to four counts each in August 2025 and were sentenced to one year of probation in December 2025. The 2014 Brendan Eich resignation following the Proposition 8 contributor disclosure [7] illustrates the donor-employer linkage harm at the state level.

D. Where this work fits

Our framing is dual to the canonical Sweeney framing. Where the Sweeney work asks "given a target individual or anonymized record, can we re-identify them by joining to a voter file?", we ask "given a voter file, what universe of adversaries can re-identify which subgroups, with what knowledge K, what budget β, what capabilities C, with what documented harm π?" The two framings are complementary; both motivate concrete reforms. The field-level Sweeney framing motivates redaction, which produces the HIPAA Safe Harbor pattern. Our framing motivates access-control reforms (rate limits, requester verification, downstream-resale prohibitions), which we argue have higher marginal privacy value at the current state of the art.

IV. Formal Framework

This section defines the analytical objects used throughout the rest of the paper and states three theoretical properties of the linkage ladder. Each definition is followed by a plain-language interpretation, and each formal result is followed by a short proof and an interpretive paragraph. The objective is to give each empirical claim in §VII a precise meaning that another researcher can reproduce exactly, and to ground the methodology in properties that hold for any voter file (or any public personally-identifying-information dataset to which the methodology is applied).

A. Notation and definitions

Definition 1. Dataset. A dataset D is a collection of N records, each describing one voter. Each record assigns a value to each of a finite set of attributes (name, address, gender, registration date, etc.). We treat the published voter file, after de-duplication on the voter unique-identifier field, as the canonical D. In plain terms: D is "the file", the rows we analyze.

Definition 2. Quasi-identifier set Q. The quasi-identifier set Q is the subset of attributes that may permit re-identification when combined and joined to external data. For voter-file analysis we take Q to include name (first, middle, last, suffix), residential address (block, street, unit, city, ZIP, ZIP+4), gender, date of birth, race, ethnicity, party, phone number, and registration date, the union of fields published anywhere across the U.S. disclosure spectrum. Different jurisdictions publish different subsets of Q. In plain terms: Q is the menu of fields that an adversary might know about a target.

Definition 3. Equivalence-class function E_F. For any subset F of Q, the equivalence class of a record r under F is the set of records in D that match r exactly on the attributes in F. We write E_F(r) for this set, and k_F(r) for its size. In plain terms: E_F(r) is the set of voters who are indistinguishable from r given only the values of F. If the attacker knows only ZIP and gender, two voters with the same ZIP and gender are in the same equivalence class.

Definition 4. k-anonymity profile. The k-anonymity profile of D under F is the distribution of k_F values across the records of D. We summarize this distribution by two numbers: u_F, the share of records that are unique under F (i.e., k_F = 1, meaning no other voter matches them on F); and rho_F, the share of records in groups of fewer than five (k_F < 5). In plain terms: u_F is the percentage of voters that an attacker who knows F can pin down to a single record. rho_F is the percentage that the attacker can narrow down to four candidates or fewer.

Definition 5. Linkage ladder L_D. The linkage ladder of D is the function that assigns to each plausible attacker knowledge set F the pair (u_F, rho_F). The ladder characterizes the inherent re-identifiability of D under every plausible adversary input; it is independent of any specific external dataset that might supply F. In plain terms: the ladder answers, for each combination of fields the attacker might know about you, how often that knowledge is sufficient to identify exactly one voter in the file. Tables II and III report the ladder for our two case files.

Definition 6. Quasi-identifier substitution. Two disjoint subsets F and F-prime of Q substitute for each other if combining either with the rest of the adversary's knowledge produces approximately the same uniqueness rate. In plain terms: removing one quasi-identifier from a published file is wasted effort if a different field of comparable identifying power remains. The empirical content of §VII-E is that registration date in Texas substitutes for date of birth: ZIP plus gender plus single-day registration date yields u = 0.2789, while ZIP plus gender plus registration year yields u = 0.0005, a ratio of 558. Texas withholds DOB but publishes registration date at single-day resolution, so the privacy benefit of the DOB redaction is largely offset.

Definition 7. Behavioral fingerprint F_v. For each voter we observe a binary turnout vector v(r) over T past elections, one bit per election, equal to 1 if the voter participated. The behavioral-fingerprint quasi-identifier is the entire vector v, treated as a single high-cardinality attribute. In plain terms: a voter's 30-year participation pattern is itself an identifier, with no need for name or address. §VII-D shows this empirically.

B. Theoretical properties of the linkage ladder

The empirical linkage ladders we report in §VII have three theoretical properties that are useful both for interpreting the empirical numbers and for reasoning about the methodology when applied to new datasets. We state them as a lemma, a composition identity, and a proposition relating the linkage ladder to the collision probability of the quasi-identifier projection. Proofs are short and direct from the definitions.

Lemma 1. *Monotonicity of the linkage ladder. For any subsets F and F-prime of the quasi-identifier set Q with F a subset of F-prime, the unique-record share satisfies u_F is at most u_{F-prime}, and similarly the small-group share satisfies rho_F is at most rho_{F-prime}.*

Proof. Fix any record r in D. Any record r-prime that matches r on F-prime must also match r on F, since F is contained in F-prime. Hence E_{F-prime}(r) is contained in E_F(r) for every r, and therefore k_{F-prime}(r) is at most k_F(r). The event "r is unique under F-prime", that is, k_{F-prime}(r) = 1, implies the event k_{F-prime}(r) is at most some constant whenever the constant is at least one. In particular, if r is unique under F (so k_F(r) = 1), then k_{F-prime}(r) is at most one and is at least one (every record matches itself), so r is unique under F-prime. The set of voters unique under F is therefore a subset of those unique under F-prime, giving u_F at most u_{F-prime}. The argument for rho is identical. □

Interpretation: adding any quasi-identifier to the adversary's knowledge cannot decrease re-identification rates. The ladder is monotone with respect to set inclusion. This is the formal

counterpart of the intuition that "more knowledge is at least as helpful for re-identification."

Lemma 2. *Composition of equivalence classes. For any disjoint subsets F_1 and F_2 of Q, the equivalence class of a record r under the union F_1 union F_2 equals the intersection of its equivalence classes under F_1 and F_2 individually: E_{F_1 union F_2}(r) equals E_{F_1}(r) intersected with E_{F_2}(r). Consequently, k_{F_1 union F_2}(r) is at most the minimum of k_{F_1}(r) and k_{F_2}(r).*

Proof. A record r-prime is in the union-equivalence class if and only if it agrees with r on F_1 union F_2, which holds if and only if it agrees with r on F_1 and on F_2 separately. The size inequality follows immediately because an intersection of two sets has cardinality at most the minimum of the two cardinalities. □

Interpretation: chaining quasi-identifier knowledge is intersective. The lemma is the formal basis for the marginal-information-gain analysis in §VII-G: each additional dataset linked into the chain corresponds to adding another quasi-identifier subset, and the equivalence class shrinks (or stays the same) with each addition.

Proposition 1. *Collision-probability bound on uniqueness. For any quasi-identifier subset F of Q, define the collision probability p_F = (1/N²) · sum over equivalence classes of m_i squared, where m_i is the size of the i-th class. Then u_F is at least 2 minus N times p_F. The bound is tight when no equivalence class has size greater than two.*

Proof. Decompose the sum of squared class sizes: sum of m_i squared equals sum over singletons of m_i squared (each contributing 1) plus sum over non-singleton classes of m_i squared. The first sum equals N times u_F, since the number of singleton classes is N times u_F. For the second sum, every non-singleton class has m_i at least 2, so m_i squared is at least 2 times m_i, and summing over non-singleton classes gives at least 2 times the count of records in non-singleton classes, which is 2 times N times (1 minus u_F). Combining, sum of m_i squared is at least N times u_F plus 2 times N times (1 minus u_F), which equals N times (2 minus u_F). Dividing by N squared gives p_F is at least (2 minus u_F) divided by N. Rearranging yields u_F is at least 2 minus N times p_F. Equality requires every non-singleton class to have m_i squared equal to 2 times m_i, i.e., m_i = 2. □

Interpretation: high collision probability of the F-projection (many records sharing the same F-values) implies low uniqueness; low collision probability implies high uniqueness. The Renyi-2 collision entropy H_2(F) = minus log of p_F is therefore a sufficient summary statistic for identifiability under F, in the sense that any two distributions of equivalence-class sizes with the same H_2(F) yield uniqueness rates within the bound above. Empirically, in §VII-B we observe u_F at 0.9581 for F equal to first name plus last name plus ZIP on the Travis file; this implies p_F is at most approximately 1.0419 / N, i.e., collision probability slightly above 1/N. The empirical collision rate is consistent with quasi-identifier values that are nearly all unique in the file.

V. Adversary Model

We now define a canonical adversary parameterization that subsumes the threat-model case studies in §X and the linkage map in §IX. The model is deliberately compact: a four-tuple capturing what the adversary knows, what they can spend, what operations they can perform, and what harm they intend.

Definition 8. Adversary. An adversary A is a tuple (K, beta, C, pi) where: K is the knowledge set, the values of certain quasi-identifiers that the adversary already possesses for the target prior to consulting D; beta is the budget, measured in dollars per target dossier; C is the capability set, drawn from {passive_query, active_query, fuzzy_match, dataset_purchase, OSINT, breach_corpus}; and pi is the harm intent, drawn from {physical, financial, reputational, discriminatory, civic, aggregate}. In plain terms: K is what the attacker already knows; beta is what they will spend; C is what they can do; pi is the kind of harm they want to cause.

Definition 9. Adversary success. Against target voter t and dataset D, a passive adversary succeeds if the attacker's pre-existing knowledge K already pins t to a single record in D, that is, if k_K(t) = 1. An active adversary additionally has access to disambiguation queries (asking neighbors, sending test mail, scraping social media) and succeeds if any sequence of such queries within budget reduces the candidate set to one. In plain terms: a passive adversary succeeds if the file alone is sufficient; an active adversary can spend additional effort to narrow ambiguities.

Definition 10. Vulnerable class. Given an adversary A, the vulnerable class V_theta(A) is the subset of records in D that fall in groups of size theta or smaller given the adversary's knowledge K. We report two key ratios: V_1, the uniquely-vulnerable share (records identifiable to exactly one voter), and V_5, the small-group-vulnerable share (records narrowed to four candidates or fewer). In plain terms: |V_1(A)| is the number of voters that the attacker can identify with certainty; |V_5(A)| is the number they can narrow to a manageable shortlist.

A. Adversary classification

We organize the threat-model case studies in §X by adversary capability tier. Tier-0 adversaries know only name and city and can only read the public file passively; this includes intimate-partner abusers, ordinary stalkers, and post-Dobbs harassers (TM1). Tier-1 adversaries additionally know the target's ZIP and possibly an inferred employer; this includes hiring managers in TM2. Tier-2 adversaries have OSINT capability and modest budget (around $30 per target); this includes most physical-safety threat actors (TM5). Tier-3 adversaries have dataset-purchase capability and larger budget (around $500 per target); this includes the asset-hunt and identity-fraud cases (TM6, TM7). Tier-4 adversaries are state-level, with resources

to combine voter files, social-media OSINT, and compromised data; this includes the deployed-military OPSEC case (TM4) and the most sophisticated suppression-mailing operators (TM3 at scale).

Costs are estimated rather than measured. We use the convention beta = $0 for tiers 0 and 1, $30 for tier 2, $500 for tier 3, and effectively unbounded for tier 4. We present empirical cost ranges in §VII-G.

VI. Methodology

A. Datasets

We use two empirical case files. Travis County, TX is a snapshot dated September 6, 2023, with 879,827 records and 372 columns; it was obtained under Texas Election Code §13.004 [11]. Robeson County, NC is a snapshot with 93,232 records (63,435 active) and 70 columns; it was obtained under North Carolina General Statutes §163-82.10 [13].

B. Step 1, Linkage map

We catalog candidate external datasets that share quasi-identifiers with the voter file. For each, we record the dataset class, the join keys it provides, the new attributes revealed when merged, the access cost (free, public-information request, commercial, or dark-web), and verified prior-art harms. The map is built by enumerating, for each quasi-identifier in the voter file, every public or commercial dataset that uses that identifier. The result is Table V (§IX).

C. Step 2, Linkage ladder

For each plausible attacker knowledge set corresponding to an entry in the linkage map, we compute the linkage-ladder values: the share of voters that are uniquely identified (k=1) and the share in groups of fewer than five (k<5). We tested 13 knowledge sets for the Travis file and 12 for the Robeson file. The combinations are derived from what each candidate in the linkage map actually publishes; we do not test arbitrary tuples.

D. Step 3, Marginal information gain

We tabulate the marginal set of attributes added at each step of a realistic chain of linkages, with rough time and dollar cost per stage. The output is Table IV (§VII-G).

E. Step 4, Harm taxonomy and case studies

We classify each linkage by its primary harm modality. Each modality is illustrated by a threat-model case study in §X, instantiated with a specific adversary class, an attack pipeline, and a vulnerable-class size computed directly from the voter file. Each case study cites at least one verified prior-art incident from litigation, journalism, or government reporting.

F. Reproducibility and ethics

All record-level analyses were performed in-memory; no individual record has been written to disk in identifiable form, displayed, or exported. A single Python script reproduce.py emits a results.json containing every numeric claim. The empirical FEC linkage in §VII-F uses an aggregate-only output script. The voter files were lawfully obtained as public records under their respective state statutes; the research targets disclosure policy, not any individual voter.

VII. Empirical Results

A. Disclosure regimes compared

Table I compares field-level disclosure between the two anchor regimes. A second-order privacy-relevant difference is the handling of Address Confidentiality Program (ACP) records: the Robeson file appears to filter ACP participants out before release (the confidential_ind column reads N for every record); the Travis file uses suspense codes that do not clearly distinguish ACP participants.

TABLE I. Disclosure regime comparison

Field	TX	NC
Full name	Yes	Yes
Residential address	Yes	Yes
Gender	90% fill	100%
Date of birth	No	Year only
Race / ethnicity	No	Yes
Phone number	No	61% fill
Party registration	Inferred	Declared
30+ year vote history	Yes	Limited
ACP records	Suspense codes	Pre-filtered

B. Linkage ladder, Travis County (TX)

Table II reports the share of Travis voters identifiable under each of thirteen plausible attacker knowledge sets. The headline result is that an attacker who knows only first name, last name, and ZIP code uniquely identifies 95.81% of voters (Wilson 95% CI [95.77, 95.85]) and narrows another 4% to groups of fewer than five. Even the inverse attack, in which the attacker already knows the residential address and observes the resident's gender, identifies 67.51% of voters uniquely. Knowing nothing but first and last name, county-wide, no geographic narrowing, uniquely identifies 75.14%.

TABLE II. Travis County linkage ladder (k=1 and k<5 percent)

Attacker knows	k=1	k<5
First, last, ZIP	95.81	99.86
Last, middle, ZIP	87.67	96.44
First, last, city	80.92	94.38
First and last (county-wide)	75.14	91.51
Last, first initial, ZIP	68.58	90.58
Address + gender (inverse)	67.51	98.71
Last, building (no unit)	58.33	98.38
Precinct, last name	39.25	79.65

ZIP, gender, exact reg-date	27.89	64.47
Full address (with unit)	26.51	92.42
Building only (no unit)	8.89	63.87
Initials + ZIP + gender	1.73	8.11
ZIP, gender, registration year	0.05	0.26

C. Linkage ladder, Robeson County (NC)

Table III reports the parallel ladder for the Robeson active-voter file. The recreation of Sweeney's classic 87% finding, combining ZIP, gender, birth year, race, party, and name initials, reaches 93.27%, very close to Sweeney's original number. The single most striking finding is that the published phone number is, on its own, a near-primary key into the file: 88.53% of voters who have a phone number listed have a number that is unique within the county. Any external dataset containing phone numbers, commercial brokers, leaked breach corpora, telemarketing lists, joins into the NC voter file at this rate without any other identifier. This is the largest single privacy delta between the two regimes; Texas has no equivalent attack because no telephone field is published.

A small-cell concern in Robeson is the county's heavily structured racial composition (American Indian, predominantly Lumbee, 32.78%; White 27.00%; Black 23.21%). 16 cells defined by ZIP-and-race contain fewer than five voters each (35 voters total); 52 cells defined by ZIP, race, and gender contain fewer than five each (115 voters total). In racially heterogeneous small counties, intersectional cells are individually small, and re-identification risk is concentrated in those cells.

TABLE III. ROBESON COUNTY LINKAGE LADDER (K=1 AND K<5 PERCENT)

Attacker knows	k=1	k<5
Phone (of phone-having voters)	88.53	,
Full Sweeney recreation	93.27	99.99
First, last, ZIP	87.79	98.61
First, last, city	86.35	98.11
First and last name	73.42	91.57
Address + gender (inverse)	67.71	99.44
Address only	29.66	94.24
Precinct, last name	16.99	41.91
ZIP, gender, birth-yr, race, party	10.06	32.66
ZIP, gender, birth-yr, race	3.83	14.14
Initials + ZIP	1.67	7.84
ZIP, gender, birth year	0.40	2.02

D. Behavioral fingerprint as identifier

The Travis voter file records turnout for 105 elections back to 1990. We treat each voter's participation pattern as a 105-bit binary string and ask whether the pattern itself is identifying. It is. Across the entire file of 879,827 voters, 24.84% have a unique turnout pattern, distinguishable from every other voter on the basis of when they voted alone, with no name, no address, no demographic data. Restricting to voters with at least 5 elections of participation, the unique share rises to 60.44% (out of 356,228 such voters). For 10 or more elections it is 86.68% (out of 216,165). For 20 or more elections it is 98.42% (out of 103,668). Long-term active voters are essentially fingerprinted by their participation history. The result generalizes beyond voter files: any longitudinal binary-participation panel, transit ridership, prescription pickup, location traces, any record of who showed up when, exhibits the same monotone-uniqueness property over time.

E. Quasi-identifier substitution: registration date as a DOB proxy

This subsection elaborates the EDRDAT finding flagged in Definition 6. Texas withholds date of birth, the field most associated with the Sweeney 87% result, but it publishes the registration date (EDRDAT) at single-day resolution. We tested whether EDRDAT carries comparable identifying power. With ZIP and gender as additional knowledge, exact registration date uniquely identifies 27.89% of Travis voters and narrows another 36.6% to groups of fewer than five. Generalizing the same field to year resolution drops the unique-identification rate to 0.05%, a 558-fold reduction in identifying power.

This is the operative reason Texas's field-level redaction does not produce the privacy benefits commonly attributed to it: the entropy that DOB would have contributed is largely contributed by the registration date instead. The corresponding policy intervention is a one-line change at file-export time, round registration date to year, with no loss of utility for any legitimate research or audit purpose. It is the cleanest narrow recommendation of the study.

F. Empirical linkage demonstration: FEC × Travis voter file

To validate the theoretical re-identification rates against a real linkage, we executed a small but realistic merge between the Federal Election Commission's individual-contribution data and the Travis voter file. We pulled 500 contribution records for ZIP 78704 (an Austin-core ZIP including South Congress and Travis Heights neighborhoods) from the 2024 cycle via the FEC OpenAPI on May 1, 2026. We de-duplicated to 181 unique contributors by exact match on (last name, first name, ZIP), and inner-joined to the voter file on the same key, no fuzzy matching, no nickname normalization, no suffix handling. Of the 181 contributors, 105 (58.01%) matched any voter record and 95 (52.49%) matched a uniquely-identifiable voter. Of the 105 matches, 74.3% had a non-trivial employer field in FEC.

The 42% unmatched share is attributable to nickname mismatches (Bob versus Robert, Beth versus Elizabeth), residential moves between the donation date and the September 2023 voter-file snapshot, hyphenation and suffix handling, and contributors registered in adjacent ZIPs but with mailing ZIP 78704 on FEC filings. Standard linkage tooling, phonetic

matching, nickname expansion, suffix stripping, routinely raises match rates above 90% in the linkage literature. The 52.49% reported here is therefore the empirical floor, not the ceiling. The theoretical 95.81% k=1 rate from §VII-B is the ceiling that a motivated attacker with conventional tooling would approach. No individual record from the merge is included in this paper; the linkage was performed in memory and only the aggregate statistics above were retained.

G. Marginal information gain through chaining

Re-identification within the file is the first step. The harm scales when the merged record is chained against additional public and commercial datasets. Table IV tabulates the marginal attributes added at each step of a realistic chain, with rough budget per target. Each row adds attributes that, individually, are high-stakes: cell phone, mortgage amount, criminal-court history, employer, photograph.

TABLE IV. MARGINAL INFORMATION GAIN THROUGH CHAINING

Stage	New attributes added	Budget
0: voter	Name; address; gender; reg-date; precinct; districts; turnout; primary signal (TX) / declared party (NC); race, byr, phone (NC only)	$0
1: people-search	Cell phone; email; age; prior addresses; relatives; aliases	$0 to 30
2: property	Assessed value; year acquired; owner; homestead; mortgage	$0-PIA
3: court	Civil suits; divorces; custody; criminal; evictions	$0
4: social	Photo; employer; network graph; location pattern	$0
5: breach	Emails; passwords; service usage; partial SSN/DL	$0 to 500
6: premium	Modeled income; financial scores; modeled race/party	$$$ enterprise

A complete dossier on a named target through Stage 4 takes 30 to 90 minutes and costs $0 to $30 total. The remaining stages add modeled-attribute information at higher cost. For comparison, legal-defense costs after a single privacy compromise, defamation, harassment, identity theft, or stalking-related litigation, routinely exceed $10,000 per incident. The asymmetry between attacker cost and defender cost is large.

H. Notable subgroup statistics

Four subgroup statistics on D_TX merit recording. (i) 4,308 voters (0.49%) carry non-Texas mailing addresses; top destinations CA (625), NY (343), VA (295). (ii) 320 voters carry military APO/FPO codes (210 AE, 102 AP, 8 AA); each is uniquely identifiable. (iii) 67,829 voters (7.71%) carry a suspense indicator. (iv) On primary-derived partisanship, 354,210 voters (40.26%) have ever pulled a primary ballot; 243,220 are lifetime DEM-only and 63,481 are lifetime REP-only, so 306,701 voters (34.86%) carry a clean partisan signal. By comparison, NC declared-party data exposes 37,919 active voters (59.78%), nearly 2×, to the same TM2 (workplace-political-screening) attack despite essentially identical name+ZIP linkability. This is a quantifiable privacy advantage of open primaries that is not visible from the file structure alone.

VIII. ROBUSTNESS ANALYSIS

A reviewer skeptical of the headline numbers will ask: Could the 95.81% finding be an artifact of how we defined uniqueness? Of which ZIPs we happened to look at? Of name-normalization choices? Of data-quality issues that we did not address? This section reports four robustness analyses that address each of these concerns in turn. The conclusion is that the point estimates are stable: tight confidence intervals, low geographic variance, small dependence on normalization choices, and counter-intuitive insensitivity to character-level noise.

A. Confidence intervals on headline statistics

We treat each k=1 share as a binomial proportion over the file and compute Wilson score 95% confidence intervals. For the headline name+ZIP attack on Travis (n = 879,827; point estimate 95.81%), the interval is [95.77%, 95.85%]. For name-only it is [75.05%, 75.23%], around a point estimate of 75.14%. For name+city it is [80.84%, 81.00%], around 80.92%. The intervals are narrow because N is large; the point estimates are not sampling artifacts.

B. Per-ZIP variation

A reviewer might worry that the headline is driven by a few large ZIPs with diverse populations, masking ZIPs where uniqueness is much lower. We tested this directly by computing the within-ZIP name uniqueness rate, the share of voters that name alone uniquely identifies, restricted to each individual ZIP. Across the 58 Travis ZIPs with at least 100 voters, the within-ZIP uniqueness has median 97.12%, interquartile range [96.17%, 97.87%], and full range [91.18%, 99.41%]. Even the worst-case ZIP exceeds 91%. Geographic skew does not materially alter the conclusion: every reasonably populated ZIP is highly name-discriminating.

C. Sensitivity to name normalization

The point estimates use case-folded, whitespace-stripped names. We re-evaluated the name+ZIP uniqueness under three additional normalization variants: (i) regex-based suffix removal of Jr, Sr, II, III, IV (which changed 200 records); (ii) Unicode NFKD accent folding (which changed records carrying tildes, accents, and ñ characters); and (iii) a 25-entry first-name nickname expansion mapping common short forms

to canonical roots, Bob → Robert, Liz → Elizabeth, Jim → James, and so on (which changed 18,061 records). The four variants gave uniqueness rates of 95.810%, 95.802%, 95.802%, and 95.534%. The maximum departure is 0.276 percentage points. The headline is robust to typical name-normalization choices.

D. Sensitivity to data perturbation

A reviewer might worry that real-world data errors, typos, OCR failures, transcription differences between voter file and external dataset, undermine the attack in practice. We modeled this by injecting independent character-level errors into name fields at rates of 1%, 5%, and 10%. Counter-intuitively, perturbation increases uniqueness: 95.81% baseline → 95.97% at 1% errors → 96.50% at 5% → 97.11% at 10%. The mechanism is that perturbation breaks colliding name-and-ZIP tuples; when a multi-voter group has one of its members' names corrupted, the group becomes singletons. So noise amplifies rather than attenuates re-identifiability. The practical consequence is that the reported point estimates are conservative under realistic data-quality assumptions: real-world errors make the attack slightly easier, not harder.

IX. Linkage Map

Table V enumerates 15 candidate dataset classes for D_TX, the join keys F each provides, the access cost, and the primary harm modality. The map is approximately representative; a researcher applying the methodology to another jurisdiction should produce a comparable artifact.

TABLE V. Linkage map (Travis County, TX)

Dataset class	Join keys F	Access	Primary harm
FEC contributions	Name+ZIP	Free bulk	Reputational, civic
TX Ethics Commission	Name+ZIP	Free	Reputational, civic
TCAD property records	Address	PIA	Financial, physical
County Clerk deeds/mortgages	Name+addr	Free online	Financial
District Court records	Name	Free online	Reputational, discrim.
Sex offender registry	Name+addr	Free	Physical
Professional licensing (TDLR, BON)	Full name	Free	Reputational, physical
Vehicle registration (DPS)	Name+addr	PIA / brokers	Physical
Voter-file vendors (L2, etc.)	VUIDNO / N+Z	$$$ wholesale	Civic, discrim.
Commercial people-search	Name+city	Free → $30/mo	Physical, reputational
Premium brokers (LexisNexis)	Name+ZIP+DOB	$$$ enterprise	Discrim., financial
Social media (LinkedIn, FB, X)	Name+city	Free	Physical, reputational
Breach corpora	Email/phone	Free → dark-web	Financial
USPS NCOA	Address	Restricted commercial	Physical, financial
HIPAA Safe-Harbor health files	ZIP+age+gender	Free	Discrim., reputational

X. Threat-Model Case Studies

Each of the seven case studies that follow describes a specific class of adversary, the steps they would take to exploit the file, the size of the affected voter population computed directly from the data, the realistic time and dollar cost of the attack, a verified prior-art incident from litigation or journalism, and the realism of available defenses. Cases are ordered by harm modality, beginning with physical-safety cases and proceeding through discrimination, civic harm, national-security, and financial harm.

A. TM1, Targeted location finding (physical-safety harm)

Adversary class: an individual who already knows the target's name and approximate city. Capability requirements are trivial, a publicly downloadable file and a CSV reader. Motivations include physical violence, harassment, and retaliation. The adversary class is broad and includes intimate-partner abusers, anti-abortion harassers post-Dobbs, anti-LGBTQ harassers, anti-trans harassers, victim-of-crime harassers, and ordinary stalkers. We classify this as a Tier-0 adversary because the knowledge set (name and city) and the budget (zero) are both minimal.

Attack pipeline: download D → filter by K → if $|E\_K(t)| > 1$, disambiguate using auxiliary observations (registration era, neighborhood signal, social-media-disclosed birth year) → retrieve full residential address with unit, precinct, districts.

Vulnerable class. On the Travis voter file, 711,968 voters (80.92% of the file) are uniquely identified by name + city alone, and 830,355 (94.38%) fall into groups of fewer than five candidates. On the Robeson active-voter file, 54,773 voters (86.35%) and 62,234 (98.11%) respectively. Time: minutes. Cost: zero. The TX file is obtainable for a nominal handling fee from the county; the NC file is freely downloadable from the state.

Prior art. Amy Boyer, October 1999: Liam Youens paid Docusearch $45 for an SSN and $109 for a work address; Docusearch obtained the work address by pretexting Boyer by telephone. Remsburg v. Docusearch (NH 2003) [6] established broker duty of care but did not modify voter-file disclosure. Post-Dobbs (2022): threats against reproductive-care workers rose 20%, stalking incidents rose 229% [8], with doxxing, including via voter rolls, among the primary modalities.

Defense. Texas operates an Address Confidentiality Program under Texas Code of Criminal Procedure Chapter 58, Subchapter B [12]; North Carolina operates one under N.C. General Statutes Chapter 15C [14]. Both are opt-in and require an affidavit or court order. Robeson appears to filter ACP participants out of the published voter file entirely; the Travis file uses suspense codes that do not clearly distinguish ACP records. The NC filter-rather-than-flag pattern is preferable; both ACPs depend on parallel filtering by every downstream agency that touches voter data, which not all agencies implement.

B. TM2, Workplace political screening (discriminatory harm)

Adversary class: a hiring manager, HR investigator, or politically motivated employer who has the applicant's name and ZIP from a resume, application, or LinkedIn profile. Capability is trivial. Tier-1 adversary.

Pipeline: take the applicant's name and ZIP from the application, look them up in the voter file, read their primary-ballot history (TX) or declared party (NC), infer or read partisan classification, integrate into the hiring decision.

Vulnerable class. In Texas, 306,701 voters (34.86% of the file) have a "clean" partisan signal, they have voted lifetime-only in DEM or REP primaries, and can be classified by reading their voting history. In North Carolina, 37,919 active voters (59.78%) declare DEM or REP party affiliation at registration and can be classified directly. The cross-state ratio is approximately 1.71: NC exposes nearly twice as many voters to this attack as TX, despite both files producing the same name+ZIP identification rate. This is the empirical content of the under-recognized open-primary privacy advantage.

Prior art. The post-2008 release of California's Proposition 8 contributor list [7], cross-referenced by activists, surfaced 1,300+ employees of 37 companies as donors and contributed to the 2014 resignation of Brendan Eich as CEO of Mozilla over a six-year-old $1,000 donation.

Defense. Federal anti-discrimination statutes (Title VII, ADA, ADEA) do not protect political affiliation. ~20 states and municipalities prohibit such discrimination under state-level statutes; enforcement requires the affected employee to know they were screened, which the voter-file inference makes nearly impossible.

C. TM3, Voter intimidation mailings (civic harm)

Adversary class: a campaign operative or politically motivated actor seeking to depress turnout among an opponent's base. Capability requires bulk file acquisition (free or near-free in most states) and mailing or robo-call infrastructure. Budget scales with target list at roughly $0.10 per recipient. Tier-2 to Tier-3 adversary.

Pipeline: bulk acquire the voter file, segment by inferred partisan lean, turnout history, and tenure of registration, mail or robo-call false election information targeted at the resulting subgroup.

Vulnerable class. The most exposed subgroup is new registrants: 79,649 Travis voters (9.05%) registered within the 12 months before the file snapshot. These voters lack prior-election experience to anchor against false claims like "your registration is invalid; you must re-register at this URL by such-and-such date." A secondary vulnerable class is the 525,617 voters (59.74%) who have never pulled a primary ballot, who cannot self-verify against past primary participation when receiving a "your party affiliation has changed" message.

Prior art. The Burkman and Wohl 2020 robocall scheme [10]: approximately 12,000 Detroit residents were called with false claims that mail-in voting placed personal information in police-accessible warrant databases; the broader campaign targeted approximately 85,000 voters across Michigan, Illinois, New York, Pennsylvania, and Ohio. The Michigan Court of Appeals affirmed the trial court's denial of the defendants' motion to quash in December 2024; the Michigan Supreme Court declined to hear an appeal in June 2025; both defendants pleaded no contest to four counts each on August 1, 2025, and were sentenced to one year of probation on December 1, 2025. The roughly five-year offense-to-resolution gap illustrates the limits of after-the-fact prosecution as a deterrent against suppression campaigns timed to a specific election cycle.

Defense. Voter-file access controls (rate limits, requester verification, audit logs, downstream-resale prohibitions) would intervene before the harm occurs but are absent in both states.

D. TM4, Deployed-military OPSEC (national-security harm)

Adversary class: a foreign intelligence service or sophisticated criminal organization seeking to identify, target, or coerce currently deployed U.S. military personnel and their families. Tier-4 adversary; budget effectively unbounded.

Pipeline: download the voter file, filter for APO/FPO mailing codes (AE for Armed Forces Europe, AP for Pacific, AA for Americas), extract residential addresses and names, cross-reference with social-media OSINT to identify family members at the home address while the service member is overseas. The result exposes deployed personnel's home addresses and family identities to targeting and coercion.

Vulnerable class. 320 Travis voters carry military APO/FPO mailing codes, 210 to Europe, 102 to Pacific, 8 to Americas. All 320 are uniquely identifiable by their voter ID, name, and current home address. Because these voters are by definition currently overseas, their families are often at the published home address without the service member's in-person presence as a deterrent. The class is small in absolute terms but high-stakes individually.

Prior art. The U.S. Government Accountability Office report GAO-26-107492 (October 7, 2025) [9] documented that publicly accessible digital data, voter files among other sources, is aggregable into "digital profiles" exposing service members, their families, and senior leaders to targeting, coercion, and disruption.

Defense. Cleanest narrow intervention: APO/FPO mailing codes should trigger automatic confidentiality filtering at file-export time. Implementation cost: three string-equality checks.

Audit cost: zero (no candidate is auditing whether deployed troops are correctly registered).

E. TM5, Inverse-attack stalking (asymmetric physical-safety harm)

Adversary class: a landlord, ex-tenant, package-delivery worker, neighbor with mailbox knowledge, hotel concierge, or anyone with prior physical-address access. The defining feature is that the adversary already has the location and seeks the identity, the inverse of the more commonly studied attack direction. Tier-0 adversary.

Pipeline: take the address (which the adversary passively already has), observe the resident's gender, look up the voter file by address and gender, retrieve name plus 30-year voting history plus precinct and district codes, chain to property records, court records, and social-media OSINT.

Vulnerable class. In Travis, 593,963 voters (67.51%) are uniquely identified by full address and gender. In Robeson, 42,955 active voters (67.71%) are uniquely identified by the same. The cross-file symmetry of these numbers is striking: the absence of DOB in Texas does not mitigate this attack, because the attack does not use DOB or any field that Texas redacts. A subset at even higher risk is the 233,274 Travis voters (26.51%) who live alone at their address, single-voter buildings, for whom the address pins down the resident deterministically with no ambiguity at all.

Prior art. Amy Boyer (1999) is partly an inverse case. Post-Dobbs reproductive-care doxxing incidents frequently begin with passively-acquired addresses (clinic license-plate observation, parking-lot identification) and resolve to identity through voter-file matches.

Defense. ACPs and mailbox-only addresses (where allowed) provide partial mitigation. Voter law in both states requires residential, not mailing, address on file; the inverse attack therefore cannot be substantially mitigated without restructuring what is published. A radical option: tiered disclosure with full address available only via a logged, credentialed request channel.

F. TM6, Asset hunting (financial harm)

Adversary class: a civil litigator, judgment-debt collector, family-law adversary, or due-diligence investigator probing dual-residency or asset structures in connection with civil discovery, divorce, or judgment enforcement. Tier-2 to Tier-3 adversary.

Pipeline: filter the voter file for records where mailing state differs from residential state, these are voters with dual residency. Cross-reference county appraisal and clerk records on both ends to identify second-home addresses, mortgage amounts, and joint-ownership structures.

Vulnerable class. 4,308 Travis voters (0.49% of the file) carry non-Texas mailing addresses, with top destinations California (625), New York (343), Virginia (295), Colorado (191), and Florida (175). The class is small in absolute terms but per-voter financial stakes are high.

Prior art. Asset hunting in family-court and judgment-collection litigation routinely uses voter files; documented in legal-practice manuals; ethically permissible within civil discovery.

Defense. None at the file level. Right intervention is downstream commercial-use restriction.

G. TM7, Stale-record identity-adjacent fraud (financial harm)

Adversary class: organized identity-fraud rings, often international. Capability requires bulk file acquisition and the ability to impersonate target individuals to mailing services, government agencies, or financial institutions. Budget ranges from approximately $0 to $500 per target. Tier-2 to Tier-3 adversary.

Pipeline: filter the voter file by suspense indicator (records where mail has been returned or address is otherwise stale), produce a target list of voters who are unlikely to still be at the published address, and execute mail-interception fraud, change-of-address fraud, or voter-registration tampering against those targets.

Vulnerable class. 67,829 Travis voters (7.71%) carry a suspense indicator. The largest two suspense reasons in the data are returned-mail flags and state-OS-team-flagged records, both indicating that the published address is likely stale. These voters are attractive targets for impersonation: legitimate communications from financial institutions and governments will not reach them, but will reach an attacker who has already effected a change-of-address.

Prior art. Sweeney et al. (2017) [4] demonstrated that voter-registration tampering attacks succeed against websites in 35 states + DC using publicly available identifiers.

Defense. Signature verification and identity proofing at registration changes vary by state; a baseline national standard is overdue.

XI. Illustrative Misuse Scenarios

These scenarios are illustrative and are directly parameterized by the empirical distributions reported in §VII; they are not independent claims and should not be read as policy advocacy. The case studies in §X describe attacker classes and aggregate vulnerable populations; this section walks through five hypothetical but realistic misuse scenarios, each calibrated to a specific empirical number from §VII, to show how the linkage ladder values translate into concrete attack pipelines that an ordinarily-resourced adversary could execute today. The scenarios are deliberately hypothetical and anonymous: no individual person, place, or incident is implied.

A. Scenario 1: an estranged ex-partner finding a survivor

A woman in her thirties leaves an abusive long-term partner in another state, drives to Austin, rents a one-bedroom apartment, and quietly registers to vote at her new address. She did not enroll in the Texas Address Confidentiality Program; the enrollment process requires an affidavit and either a court order or a signed statement from a participating agency, and she has

not yet found the time or legal support to complete it. Her ex-partner, who is now released after serving time for domestic violence, knows only her first and last name. He reads online that Texas voter records are public. He spends ten minutes on the Travis County Tax Assessor-Collector's site, downloads the voter file as a CSV, opens it in a spreadsheet program, sorts by last name, and finds her in the small group of voters with her exact first and last name in Travis County. The 75.14% k=1 statistic from §VII-B says that for three out of four such targets, only one record will match, and even when there are several, narrowing by approximate age range or by which precinct he passes through to her likely workplace will resolve the ambiguity. Within the hour, he has her residential street address with apartment number, her voter precinct (which corresponds approximately to her neighborhood), and a voting history that confirms she is the same person, she had voted in primaries in their old state in years he remembers her doing so. The Travis ACP could have prevented this, but only if she had enrolled in time.

B. Scenario 2: a politically motivated employer screening applicants

A small business owner in a Texas suburb is hiring for a managerial role and has narrowed the candidate pool to four applicants. The owner is politically active and would prefer not to hire someone whose primary-ballot history suggests they belong to the opposing political coalition. Federal anti-discrimination statutes do not protect political affiliation, and Texas has no state-level prohibition that the applicants could realistically invoke. The owner takes each applicant's name from their resume, they all included ZIP codes in the contact information, and searches the voter file. By the §VII-B linkage ladder, name plus ZIP uniquely identifies the matching voter record 95.81% of the time. For two applicants, the lifetime primary-ballot history is exclusively in one party; for the third, it is exclusively in the other; the fourth has never participated in a primary and carries no signal. The owner schedules second-round interviews accordingly. None of the applicants knows why they were or were not advanced. The 34.86% of Travis voters with a clean partisan signal from §VII-A is a reasonable estimate of the share of any given applicant pool that this attack will classify; the remaining majority of applicants are simply not classifiable from voter-file data alone, but the attack does not need to classify everyone, it needs to classify those who reveal a signal.

C. Scenario 3: a foreign service mapping a deployed-military neighborhood

A foreign intelligence service is preparing an influence operation directed at U.S. military personnel currently deployed to allied bases in Europe. They want to identify, by family address, U.S. service members whose immediate family is currently at home alone in Texas, in order to target the family with disinformation, recruit them as inadvertent OSINT sources, or, in extreme cases, physically locate them. They download the Travis voter file, which is not export-controlled, and filter for voters whose mailing address carries the APO code AE (Armed Forces Europe). The §VII-H subgroup statistic identifies 210 such voters in Travis alone. Each carries a residential-address field that resolves to their family home. The operation cross-references with social-media OSINT to identify the spouses, children, parents, or roommates currently living at each address; the families have, for the most part, left a public footprint that places them at the address. The total time investment is on the order of a working day for a junior analyst. The Government Accountability Office report of October 2025 on "digital profiles" of service members [9] describes precisely this aggregation pipeline.

D. Scenario 4: a campaign operative running a suppression mailing

A political consulting firm working for an out-of-state client purchases the Travis voter file at the bulk public-records rate. They write a software tool that segments the file by tenure of registration (§VII-H: 79,649 voters, or 9.05% of the file, registered in the prior 12 months), by inferred partisan lean (the 27.6% lifetime-DEM-only and 7.2% lifetime-REP-only subgroups), and by turnout history (low-frequency voters are more vulnerable to confusion than high-frequency voters). They mail a few thousand selected new-registrant voters from the opposing party's lean group a postcard claiming that recent registrations require in-person re-verification at a specific government office by a specific date that is, in fact, after the election. A minority of recipients comply, lose their registration timing, and miss the election. The Burkman and Wohl 2020 robocall scheme [10] is the documented analog: Burkman and Wohl targeted approximately 12,000 voters in Detroit ZIP codes with similar misinformation. Both pleaded no contest to four counts each in August 2025 and were sentenced to one year of probation in December 2025, by which time the targeted election had long since concluded.

E. Scenario 5: an identity-fraud ring exploiting the suspense list

An organized identity-fraud ring acquires the Travis voter file and filters for voters with a suspense indicator (§VII-H: 67,829 voters, or 7.71% of the file). These are voters whose mail has been returned or whose addresses are otherwise stale; they are by definition unlikely to be receiving mail at the published address. The ring submits change-of-address requests to USPS for these voters, redirecting their mail to addresses the ring controls. Incoming mail from financial institutions, governments, and service providers, letters about new credit cards, statements, identity-verification codes, jury notices, flows to the ring's addresses instead of being detected as missing by the voter. The ring opens fraudulent credit accounts, intercepts identity-verification mailings, and in some cases tampers with the voter's registration record to impersonate them at the polls. Sweeney et al. [4] documented the registration-tampering vector in 2017. The financial-fraud variant has received less academic attention but has been documented in identity-theft case work.

XII. Key Findings

Finding 1. Both files admit u_{name+ZIP} > 0.85 (TX 0.9581 [Wilson 95% CI 0.9577 to 0.9585]; NC 0.8779). u_{name only} = 0.7514 (TX) and 0.7342 (NC). The disclosure-regime difference is observable but does not change the qualitative conclusion of high re-identifiability under low-knowledge adversaries.

Finding 2 (names dominate, not DOB). In neither file is date of birth the dominant identifier. Name plus any geographic narrowing produces uniqueness above 80% in both regimes; the marginal contribution of birth year over name and ZIP is small. The classic Sweeney 87% triple is recovered against the NC file (at 93.27%) only after name initials are added on top of ZIP, gender, birth year, race, and party, meaning names are doing most of the identifying work, not the demographic triple.

Finding 3 (substitution). EDRDAT substitutes for DOB on D_TX. u_{(ZIP, gender, EDRDAT)} = 0.2789 vs. u_{(ZIP, gender, EDRYR)} = 0.0005 (558× ratio). The disclosure regime that withholds DOB but publishes single-day registration date does not realize the privacy gains commonly attributed to the DOB removal.

Finding 4 (behavioral fingerprint). u_{F_v} grows monotonically with participation count: 0.2484 over all of D_TX, 0.6044 conditional on sum(v) ≥ 5, 0.8668 on sum(v) ≥ 10, 0.9842 on sum(v) ≥ 20. Long-term active voters are uniquely identifiable from participation pattern alone, with no demographic input. The result generalizes to any longitudinal binary-participation panel.

Finding 5 (inverse symmetry). u_{address+gender} is effectively constant across the two regimes (0.6751 TX, 0.6771 NC). The absence of DOB in TX does not mitigate the inverse attack class because the attack does not use DOB.

Finding 6 (open-primary advantage). TX u_{partisan signal} = 0.3486 (lifetime DEM-only or REP-only); NC u_{declared party} = 0.5978. Open primaries deliver a quantifiable workplace-screening privacy advantage of approximately 25 percentage points in the population fraction exposed to TM2.

Finding 7 (empirical floor). Real linkage of |C| = 181 unique FEC contributors in ZIP 78704 to D_TX by exact name+ZIP yields 52.49% k=1 match rate, 58.01% any match. With standard linkage tooling, the rate would approach the theoretical 0.9581 upper bound. The empirical floor confirms the theoretical ladder is not driven by data artifacts.

Finding 8 (robustness). Wilson 95% CIs on headline statistics are tight (≤ 0.10 percentage points). Across 58 ZIPs in D_TX, within-ZIP u_{name} has median 0.9712 and IQR [0.9617, 0.9787]; the worst ZIP is 0.9118. Across four name-normalization variants, u_{name+ZIP} departs by ≤ 0.276 percentage points from baseline. Counter-intuitively, character-level error injection (rates 1%, 5%, 10%) increases u_{name+ZIP} (95.81% → 95.97% → 96.50% → 97.11%) because errors break collisions. The point estimates are conservative under realistic data-quality assumptions.

Finding 9 (chained gain). A complete dossier on a named target through chain stages 0 to 4 (voter file plus people-search plus property records plus court records plus social media) takes 30 to 90 minutes at total budget under $30. Subsequent stages, breach corpora, premium brokers, add modeled-attribute information at higher cost.

Finding 10 (subgroup salience). Specific subgroup vulnerabilities on D_TX are large in absolute terms: 320 deployed-military records, 79,649 (9.05%) recent registrants, 67,829 (7.71%) suspense-list voters, 4,308 (0.49%) out-of-state mailers. These are classes of thousands of voters, not edge cases.

XIII. Discussion

A. Field-level redaction as an empirically secondary lever

Findings 1 through 3 jointly establish that field-level redaction has reached diminishing marginal returns under the disclosure regimes contemporary U.S. states currently operate. Texas, with the most conservative U.S. regime in the variables we examined, admits 95.81% unique re-identification by name and ZIP (Wilson 95% CI [95.77%, 95.85%]). North Carolina admits 87.79%. Names plus any geographic narrowing achieve 80%+ uniqueness in both files. The Sweeney triple, long the focal point of disclosure-policy attention, is empirically dominated by names plus ZIP in real voter files, and Texas's redaction of date of birth is largely offset by the registration-date substitution effect. We do not claim that no field-level intervention has marginal value; we claim that the marginal value of further field redaction is empirically smaller than the marginal value of access-control interventions that have been comparatively neglected.

The set of access-control interventions consistent with the empirical evidence comprises: (i) rate limits on bulk file requests; (ii) requester-identity verification with audited stated-purpose declarations; (iii) prohibitions on commercial resale of voter-file data, with the "publicly-available-information" carve-outs in state privacy laws revisited; (iv) audit logs of which records have been accessed by which requesters, available to affected voters via a parallel right-of-access mechanism; (v) tiered disclosure, with full record access only via a credentialed request channel and aggregated or generalized data for general public access. These interventions operate on what an adversary can do with the file, not on what is in the file.

B. Two narrow recommendations

Two specific, narrow recommendations follow directly from the data and are low-cost relative to their privacy gains. First, on the Texas voter file, generalize the registration-date field from single-day to year resolution before publication. This drops the relevant uniqueness rate from 27.89% to 0.05%, a 558-fold reduction, with no loss of utility for any legitimate research or audit purpose. The implementation is a one-line change at file-export time. Second, in both states, apply automatic confidentiality filtering to APO/FPO mailing codes (AE, AP, AA) before file export. The 320 deployed-military records in the Travis file are individually exposed to high-stakes targeting, and the implementation cost is three string-

equality checks. Both recommendations have been confirmed by the data in this study and could be implemented within a single budget cycle.

C. Note on the open-primary advantage

Finding 6, that open-primary regimes carry a quantifiable privacy advantage on TM2, does not appear in the file structure; it is a property of the underlying primary-election regime. We do not advocate for open primaries on this ground alone, but we record the privacy ledger as one item in the broader open-vs-closed debate.

XIV. Limitations

The two anchor cases were selected to bound the U.S. disclosure spectrum: Travis County represents a conservative regime that withholds DOB, race, party, and phone, and Robeson County represents a permissive regime that publishes all four. The findings reported here characterize the linkage behavior at these two anchor points. By Lemma 1 (monotonicity), any U.S. jurisdiction publishing a strict superset of the Travis fields admits at least the Travis re-identification rates; any jurisdiction publishing a subset of the Robeson fields admits at most the Robeson rates. In particular, similar linkage behavior is expected wherever name and geographic identifiers are present, which is every state in the United States. Whole-state file replication (Florida is bulk-downloadable and the natural next case) and additional county samples are nevertheless valuable for refining the per-ZIP variance characterization in §VIII-B and for stress-testing the open-primary advantage finding (§VII-H, Finding 6) on a state with closed primaries.

Several specific limitations apply. Both files are point-in-time snapshots; an adversary with multiple snapshots learns information beyond what we measured (e.g., when a voter moved or re-registered). The NC phone field is 61% populated; reported phone-uniqueness rates are conditional on field presence. The linkage map enumerates 15 dataset classes; adversaries with additional budget can purchase further commercial sources whose product offerings change over time. The empirical FEC linkage uses a single ZIP and 500 records; a larger sample with name normalization would tighten the empirical floor. We have not quantified modeled-attribute leakage, the case where a commercial broker infers a protected attribute (race, party, income) from voter-file features and re-introduces the inferred attribute into nominally-blind downstream decisions; this is a planned extension. The adversary model assumes economically-rational adversaries; sophisticated state-level adversaries (TM4) may behave outside this assumption.

XV. Conclusion

A voter file functions, in practice, as a join key into the rest of a person's public and commercial record. We formalized this role through the linkage-ladder function, the equivalence-class function, three theoretical properties (monotonicity, composition, and a collision-probability bound), and a canonical adversary model parameterized by knowledge, budget, capability, and harm intent; we applied the formalism to two voter files at opposite ends of the U.S. disclosure spectrum; we reported robustness analyses including Wilson confidence intervals, per-ZIP variance over 58 ZIPs, four normalization variants, and three perturbation rates; and we demonstrated the methodology empirically against the FEC contributions database. The two principal empirical findings, that both files admit greater than 85% re-identification under trivial adversary knowledge sets, and that the most-debated disclosure choice (date of birth) is not the dominant identifier in either file, together imply that field-level redaction is an empirically secondary lever rather than the dominant determinant of contemporary voter-file privacy outcomes. The policy frontier consistent with the empirical evidence runs through access-control interventions: rate limits, requester-identity verification, audit logging, and downstream-resale prohibitions. The methodology generalizes to any jurisdiction's voter file and, more broadly, to public personally-identifying-information datasets where name and geographic identifiers are present; planned extensions include additional state voter files, a quantification of modeled-attribute leakage, and tightening of the empirical-linkage lower bounds via standard name-normalization tooling.

XVI. References

[1] L. Sweeney, "Simple demographics often identify people uniquely," Carnegie Mellon University, Laboratory for International Data Privacy, Working Paper LIDAP-WP4 (also Data Privacy Working Paper 3), Pittsburgh, PA, 2000. Available: https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf

[2] L. Sweeney, "k-Anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557 to 570, 2002.

[3] L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the Personal Genome Project by name (a re-identification experiment)," arXiv preprint arXiv:1304.7605, April 2013.

[4] L. Sweeney, J. S. Yoo, and J. Zang, "Voter identity theft: Submitting changes to voter registrations online to disrupt elections," Technology Science, 2017090601, September 6, 2017. Available: https://techscience.org/a/2017090601/

[5] J. Sherman, "People search data brokers, stalking, and 'publicly available information' carve-outs," Lawfare, October 30, 2023. Available: https://www.lawfaremedia.org/article/people-search-data-brokers-stalking-and-publicly-available-information-carve-outs

[6] Remsburg v. Docusearch, Inc., 149 N.H. 148, 816 A.2d 1001 (2003).

[7] Mozilla Foundation, "FAQ on CEO resignation," The Mozilla Blog, April 5, 2014. Available: https://blog.mozilla.org/en/mozilla/faq-on-ceo-resignation/. See also: B. Eich, statement on resignation, April 3, 2014; coverage of the broader Proposition 8 contributor disclosure regime in The Guardian and ABC News, April 2014.

[8] 19th News, "What is doxxing? And why is a constant worry for reproductive rights workers?" April 2024.

[9] U.S. Government Accountability Office, "Information Environment: DOD Needs to Address Security Risks of Publicly Accessible Information," Report GAO-26-107492, October 7, 2025 (publicly released November 17, 2025). Available: https://www.gao.gov/products/gao-26-107492

[10] Michigan Attorney General, People v. Burkman and Wohl (Wayne County 3rd Circuit; Hon. Margaret VanHouten). Procedural history: charges filed October 2020; Michigan Court of Appeals affirmed denial of motion to quash (Dec. 2024); Michigan Supreme Court declined to hear appeal (June 30, 2025); defendants pleaded no contest to four counts

each (Aug. 1, 2025); both sentenced to one year of probation (Dec. 1, 2025). Press releases: https://www.michigan.gov/ag/news/press-releases (Aug. 2025, Dec. 2025).

[11] Texas Election Code §13.004 (voter registration record disclosure).

[12] Texas Code of Criminal Procedure Chapter 58, Subchapter B (Articles 58.051 to 58.062), Address Confidentiality Program for Victims of Family Violence, Sexual Assault, Stalking, and Trafficking, administered by the Office of the Attorney General.

[13] North Carolina General Statutes §163-82.10 (voter registration records public-disclosure provisions).

[14] North Carolina General Statutes §15C-1 et seq. (Address Confidentiality Program).

[15] HIPAA Privacy Rule, Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. parts 160 and 164, particularly §164.514(b)(2) (Safe Harbor de-identification).

[16] Federal Election Commission, OpenAPI for individual contributions, https://api.open.fec.gov/, accessed May 2026.

[17] Servicemembers Civil Relief Act, 50 U.S.C. §3901 et seq.

[18] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," Journal of the American Statistical Association, vol. 22, no. 158, pp. 209 to 212, 1927.