Appendix A. Entity Name Standardization Pipeline

This appendix documents the construction of the entity name standardization pipeline used to harmonize the names of financial and non-financial firms appearing in the collateral data. The goal of the pipeline is to group spelling variants, abbreviations, and extended legal designations under a single entity identifier and a canonical "standard name", while minimizing both false matches (distinct entities incorrectly merged) and missed matches (identical entities not merged). The design follows best practices from the record linkage and entity resolution literature (Fellegi and Sunter, 1969; Winkler, 2006; Elmagarmid et al., 2007; Christen, 2012) and adapts them to the specifics of patent-collateral data as in Mann (2017).

The pipeline consists of seven stages. First, an *exploration* phase computes descriptive statistics and inspects frequent patterns in raw names (e.g., functional roles such as "as collateral agent", legal suffixes, and geographic qualifiers). This informs the construction of rule-based normalisation dictionaries and regular expressions. Second, a *normalisation* phase applies deterministic text cleaning to each name: converting to upper case; harmonising punctuation; removing functional roles and contractual qualifiers; standardising legal suffixes (e.g., "Corporation" $\rightarrow$ "CORP", "Limited" $\rightarrow$ "LTD"); and cleaning common stopwords such as leading "THE". Such rule-based standardisation is standard in large-scale entity resolution because it concentrates variation on a reduced token space and improves the performance of subsequent similarity measures (Winkler, 2006; Christen, 2012).

Third, a *blocking* phase partitions the universe of names into candidate sets that are likely to contain true matches. Specifically, names are indexed by their first "significant" word after normalisation (skipping generic terms such as "BANK" or "COMPANY") and, for very large blocks, further sub-blocked using the first two significant words and name length. Blocking is a standard device to reduce the quadratic complexity of all-pair comparisons while preserving most true matches (Baxter et al., 2003; Papadakis et al., 2013; Christen, 2012). In this context, blocking on the first significant token is well suited because business names typically begin with a distinctive brand ("WELLS FARGO", "CREDIT SUISSE") followed by functional or legal descriptors that carry little identifying power.

Fourth, within each block the algorithm performs *fuzzy string matching* using the WRatio score from the `rapidfuzz` library, which combines character-based and token-based variants of the Levenshtein distance and related metrics (Levenshtein, 1966; Cohen et al., 2003; Winkler, 1990). Two normalised names are linked if their WRatio is at least 88 on a 0–100 scale. The threshold was calibrated empirically by examining problematic pairs and balancing the trade-off between false positives and false negatives; increasing the threshold from 85 to 88 eliminated a large share of clearly spurious matches while preserving the vast majority of

intuitive matches. This approach parallels the string similarity-based matching strategies used by Mann (2017) to harmonise collateral entity names in patent-secured lending data.

Fifth, the pairwise links obtained from fuzzy matching are interpreted as edges in an undirected graph whose nodes are individual names, and *connected components* of this graph are treated as candidate entities (Getoor and Machanavajjhala, 2012; Bhattacharya and Getoor, 2007). This transitive-closure step allows the algorithm to group chains of similar names (e.g., "WELLS FARGO BANK NA", "WELLS FARGO BANK, N.A.", "WELLS FARGO BANK NATIONAL ASSOCIATION") even when some pairs are not directly compared or fall just below the similarity threshold. Sixth, a *grouping and ID assignment* phase assigns a unique identifier to each connected component and selects a *standard name* within the component, preferring the name with the highest observed frequency and, conditional on that, the shortest variant. This choice reflects the idea that more frequent and shorter names are closer to the economically salient "brand" used in practice (Christen, 2012).

Seventh, a *validation and completion* phase flags components for review when their average similarity is low, the minimum similarity is very low, or the group is unusually large, and it then adds all remaining singletons (names without any match) as standalone entities. This step ensures complete coverage of the original universe of names, while allowing targeted manual inspection of a small subset of borderline cases, as recommended in the applied record-linkage literature (Elmagarmid et al., 2007; Christen, 2012). Overall, the combination of aggressive normalisation, conservative similarity thresholds, graph-based clustering, and explicit handling of singleton names yields a transparent and replicable mapping from raw collateral names to economic entities that is suitable for the empirical analysis of creditor rights and innovation in the main text.

References

Bhattacharya, I. and L. Getoor (2007). Collective entity resolution in relational data. *IEEE Transactions on Knowledge and Data Engineering* 19(9), 1161–1173.

Baxter, R., P. Christen, and T. Churches (2003). A comparison of fast blocking methods for record linkage. In *Proceedings of the KDD'03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation.*

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer.

Cohen, W. W., P. Ravikumar, and S. E. Fienberg (2003). A comparison of string distance

metrics for name-matching tasks. In *Proceedings of the IJCAI-03 Workshop on Information Integration on the Web*.

Elmagarmid, A. K., P. G. Ipeirotis, and V. S. Verykios (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering* 19(1), 1–16.

Fellegi, I. P. and A. B. Sunter (1969). A theory for record linkage. *Journal of the American Statistical Association* 64(328), 1183–1210.

Getoor, L. and A. Machanavajjhala (2012). Entity resolution: Tutorial. *Proceedings of the VLDB Endowment* 5(12), 2018–2019.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), 707–710.

Mann, W. (2017). Creditor rights and innovation: Evidence from patent collateral. Working paper, UCLA Anderson School of Management.

Papadakis, G., E. Ioannou, T. Palpanas, C. Niederee, and W. Nejdl (2013). A blocking framework for entity resolution in highly heterogeneous information spaces. *IEEE Transactions on Knowledge and Data Engineering* 25(12), 2665–2682.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 354–359.

Winkler, W. E. (2006). Overview of record linkage and current research directions. U.S. Census Bureau.