Friedrich-Alexander-Universität
Erlangen-Nürnberg

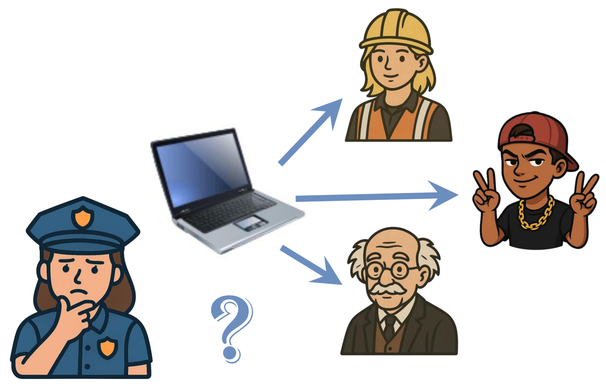
Digital Forensic Profiling with Demographic Attributes: A Research Agenda

Marion Liegl

Motivation and Idea

- Criminal law applies exclusively to humans
- The findings of digital forensic experts usually only relate to objects and accounts
- Identifying the person responsible is essential for prosecuting criminal offences
- Methods that create a profile with demographic attributes of the main person who used analyzed digital devices would be helpful

The key idea of this Phd project is to infer demographic attributes such as gender, age, native language, level of education, or occupation of the main device user based on unstructured data originating from that device.



Research Questions

- Which demographic attributes can be detected in unstructured computer data?
- Can methods developed for large datasets be applied to small, unstructured data?
- How accurate and reliable are the identified attributes?
- Which digital traces (e.g., text files, media files or technical artifacts) are most conducive for attributes identification?
- Which methods researched to date can be combined?
- To what extent can detected attributes identify an individual within a small group?
- What forms of bias influence the results, and how can this biases be detected and mitigated?

Research Steps

In order to develop initial methods for identifying useful demographic attributes, the following steps are planned:

- **Literature Review:** Existing studies address gender [1], age [2], native language [3], education [4], and occupation [5].
- **Model Development:** An initial model serves as a basis for further research (Fig. 1).
- **Prototype Development:** Implementation of methods from other disciplines for unstructured single-user system data.
- **Research Question**: The prototype will help answer our developed questions.
- **Practical Tests:** Planned evaluation with police authorities, partners and volunteers who can provide real world data.

Model

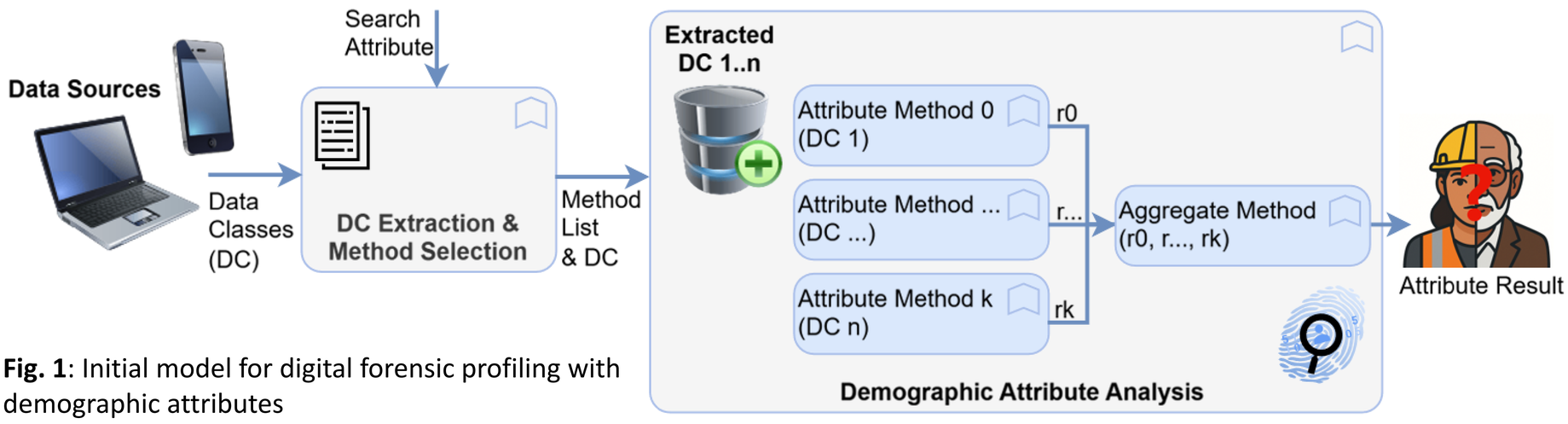


Fig. 1: Initial model for digital forensic profiling with demographic attributes

- Data from various unstructured sources (e.g., laptops, mobile devices) are analysed under controlled single-user conditions
- Relevant data classes (e.g., text, media files, browsing traces, communication, folder structures) for a specific attributes (e.g., age, gender) are extracted and processed using suitable methods - e.g., stylometry, metadata analysis, pattern recognition
- Results are aggregated to enable cross-validation, increasing reliability and reducing individual method weaknesses

References

[1] Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2018, August). Simply the best: minimalist system trumps complex models in author profiling. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 143-156). Cham: Springer International Publishing.

[2] Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007, September). Author profiling for English emails. In Proceedings of the 10th conference of the Pacific Association for computational linguistics (Vol. 263, p. 272).

[3] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

[4] Volkova, S., & Bachrach, Y. (2015). On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. Cyberpsychology, Behavior, and Social Networking, 18(12), 726-736.

[5] Jiang, Z., Yu, S., Qu, Q., Yang, M., Luo, J., & Liu, J. (2018, April). Multi-task learning for author profiling with hierarchical features. In Companion Proceedings of the The Web Conference 2018 (pp. 55-56).

Contact: marion.liegl@fau.de