United States Environmental Protection Agency

ExpoPath: Identifying and Annotating Exposure Pathways from Chemical Co-occurrence Networks

Michael A. Zurek-Ost,[1] K. Phillips,[2] A. Williams,[2] A. Edelman-Munoz,[3] S. Handa,[2] and K. Isaacs[2].

[1]Oak Ridge Institute for Science and Education, Oak Ridge, TN; [2]EPA ORD CCTE, Research Triangle Park, NC; and [3]Oak Ridge Associated Universities National Student Services Contract, Oak Ridge, TN.

Abstract #: 3338
Poster Board #: P461

Background image credit: "The Biesbosch of the Netherlands" NASA Earth Observatory image by Lauren Dauphin, using Landsat data from the U.S. Geological Survey

Michael A. Zurek-Ost I ZurekOst.Michael@epa.gov I Orcid ID: 0000-0001-5013-4240

Background and Purpose

The US EPA monitors and evaluates risk to environmental, ecological, and human health posed by chemical contaminants. Additionally, the agency analyzes confirmed and suspected modes of traversal from commercial and industrial sectors to, and across, vulnerable media and points of exposure. Understanding such relevant exposure pathways remains integral for the EPA's commitment to chemical prioritization and regulation.

Extensive work has culminated in datasets and repositories such as ChemExpo[1], a public facing dashboard synthesizing chemical commercial and functional-use data across millions of records, with regulation under the Toxic Substances Control Act (TSCA) resulting in the substantial portion of reported industrial source data, also contained in the EPA's Office of Pollution Prevention and Toxics' Chemical Data Reporting[2] (CDR) public database. Furthermore, information on chemical presence in environmental, ecological, and human systems housed within the EPA's Multimedia Monitoring Database[3] (MMDB) provides insight into chemical fate, transport, and patterns of traversal as well as eco-/bio-accumulation.

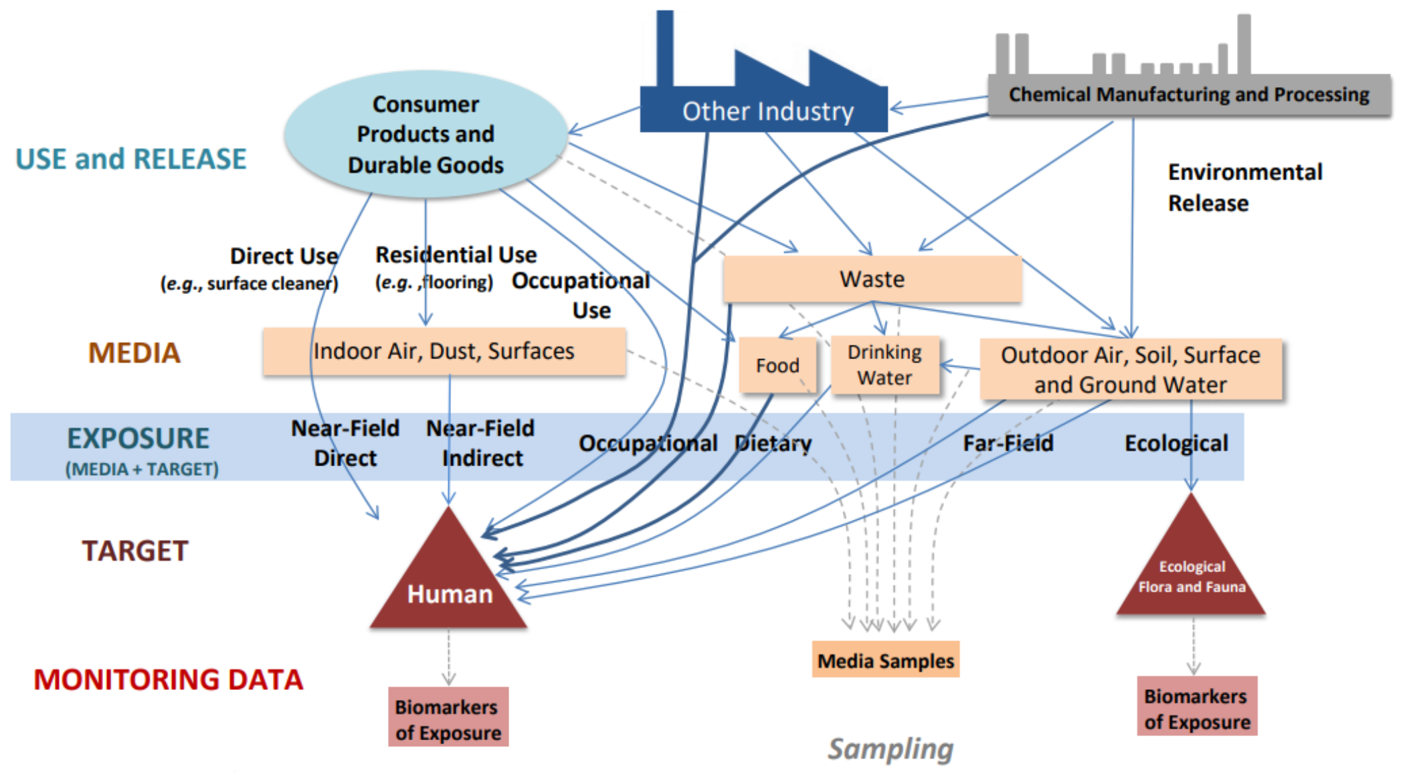


Fig 1. *A priori exposure route schematic denoting commonly-held source-to-receptor pathways.*

This research adopts a combination of network analysis and graph machine learning methodologies to deductively define likely modes of traversal between industrial and commercial sources and environmental, ecological, and physiological media/receptors derived from chemical co-occurrence patterns. This may improve upon current characterization of exposure pathways based on an *a priori* classification architecture predominately derived from subject matter expert consensus. An analysis of the patterns of chemical co-occurrence within and between source and sink media categories may reflect known source-to-receptor linkages but might also suggest a more nuanced and explicit set of exposure pathways.

Datasets

Environmental/Ecological Media
- Multimedia Monitoring Database (MMDB) → *Blood, Surface Water, Landfill Leachate…*
  - *Rapid exposure information (confirmations only)*

Consumer Products & Industrial Processes
- Chemicals and Products Database (CPDat) → *Pesticides, Pet & Animal Care, Personal Care…*
  - *Contains media in the form of consumer product categories*
- Chemical Data Reporting (CDR) → *Industrial Gas Manufacturing, Construction, Petrochemical Manufacturing…*
  - *Rule under the Toxic Substances Control Act requiring manufacturers and importers to provide information on chemical use in commerce. Contains industrial use categories*

Breakdown Products
- Chemical Transformations Database (CheT)
  - *Contains confirmed parent chemical to breakdown product relationships (see Methods - Network Formation)*

Pharmaceutical
- DrugBank → *Pharmaceuticals*
  - *Curated list of chemicals provided by University of Alberta*
- Food and Drug Administration's (FDA) Orange Book → *Pharmaceuticals*
  - *Approved Drug Products with Therapeutic Equivalence Evaluations*

Food Additive/Contact
- Food Additive and Contact Lists → *Adhesives and Components of Coatings, Paper and Paperboard Components…*
  - *Categories from FDA's Code of Federal Regulations Title 21 – Food and Drug*

Methods

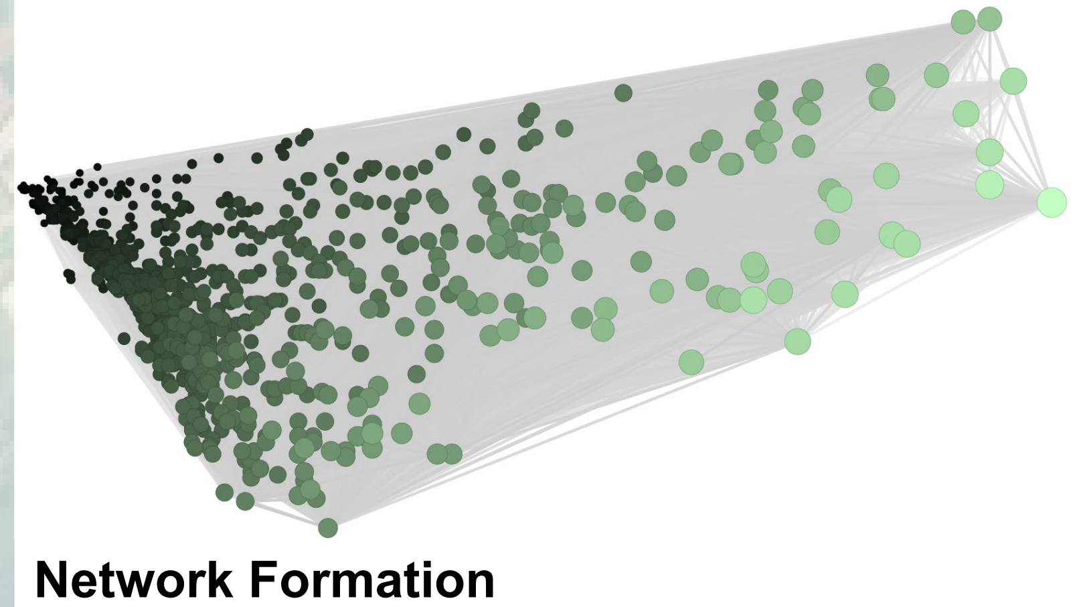


Fig 2. *One-mode chemical-to-chemical network. Node size and color reflect a chemical's PageRank[4] (algorithm developed by Google to quantify a node's importance based on connections to other well-connected nodes), which indicates how pervasive a chemical is across media categories as well as how likely such a chemical is to co-occur with other pervasive chemicals. Several chemicals co-occur with other similarly less-pervasive chemicals (left dark-green cluster). Edge color indicates both the strength and significance (t-statistic). Layout generated using Multi-Dimensional Scaling[5] (MDS) based on number of shared media between chemicals. Chemicals shown closer together appear in similar media. Nodes: 1,240. Edges: 521,461.*

Network Formation

- Media categories from the gathered datasets are designated as "source" and "sink"
- MMDB media are designated as sinks, while others represent industrial/consumer sources
- These data are filtered via three criteria:
  - Chemicals with at least one occurrence in both a source and sink are preserved
  - Inorganic compounds are removed
  - Quadratic Assignment Procedure[6] (QAP) is used to assess each relationship between any two chemicals using their respective media-by-media adjacency matrices
    - Leverages Monte Carlo Simulations to randomize one of the matrices of a pair to produce significance values and the strength of these relationships (as well as a t-statistic)
    - Once completed, insignificant edges are removed

Overlapping Community Detection (BIGCLAM)

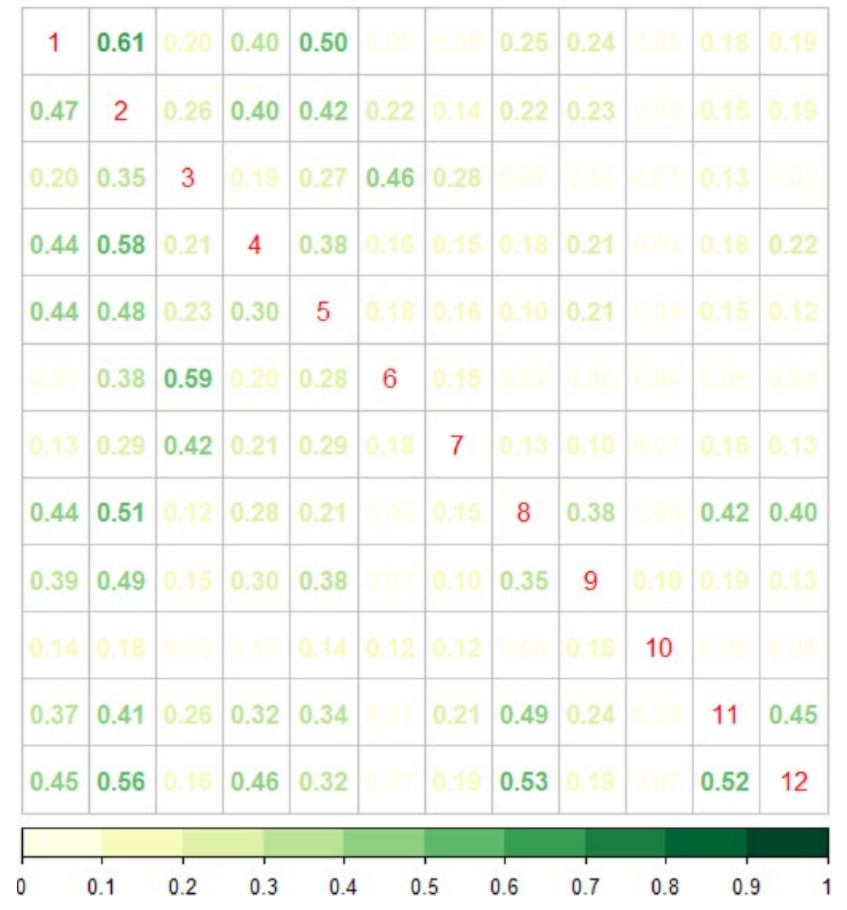
- Identifying exposure pathways from chemical co-occurrence patterns is a clustering task, where communities of related co-occurring chemicals are identified.
- Chemicals can traverse more than one exposure pathway
  - Need for an overlapping, large-scale community detection algorithm is needed
- "BIGCLAM: Cluster Affiliation graph Model for BIG networks" [7] from the Stanford Network Analysis Project[8] (SNAP)
  - turns a community detection problem into non-negative matrix factorization completed via maximum likelihood estimation. The number of communities to detect is determined by the model rather than the researchers designating a set number beforehand

Univariate Logistic Regression

To assess relationships between media, functional use categories, and physiochemical properties, with overlapping communities, separate univariate logistic regression[9] models were run for each community independently. Since the goal was to assess relationships with these variables rather than build an inclusive, explanatory and/or predictive model, a collection of univariate models were run, with their results compiled to define and characterize these communities.

Results

- 12 overlapping communities were identified using BIGCLAM
- Logistic regression results indicate differing, albeit not exclusive, relationships with a majority of source and sink categories, functional use categories, and physiochemical properties
- A highlighted example is presented in Discussion and Conclusion



1	0.61	0.20	0.40	0.50	0.05	0.08	0.25	0.24	0.05	0.18	0.19
0.47	2	0.26	0.40	0.42	0.22	0.14	0.22	0.23	0.05	0.15	0.19
0.20	0.35	3	0.19	0.27	0.46	0.28	0.07	0.10	0.01	0.13	0.07
0.44	0.58	0.21	4	0.38	0.16	0.15	0.18	0.21	0.04	0.18	0.22
0.44	0.48	0.23	0.30	5	0.18	0.16	0.10	0.21	0.04	0.15	0.12
0.07	0.38	0.59	0.20	0.28	6	0.15	0.02	0.06	0.06	0.05	0.04
0.13	0.29	0.42	0.21	0.29	0.18	7	0.13	0.10	0.07	0.16	0.13
0.44	0.51	0.12	0.28	0.21	0.03	0.15	8	0.38	0.05	0.42	0.40
0.39	0.49	0.15	0.30	0.38	0.07	0.10	0.35	9	0.10	0.19	0.13
0.14	0.18	0.03	0.10	0.14	0.12	0.12	0.06	0.18	10	0.05	0.08
0.37	0.41	0.26	0.32	0.34	0.07	0.21	0.49	0.24	0.03	11	0.45
0.45	0.56	0.16	0.46	0.32	0.07	0.19	0.53	0.19	0.02	0.52	12

Fig 3. *Percentage of overlapping chemicals between community$_i$ (row number) and community$_j$ (column number). Communities 1, 2, and 3 are large, inclusive clusters. Communities 8, 11, and 12, while smaller, share many chemicals in common with one another. Chemical overlap among communities may reflect multiple pathways chemicals are likely to traverse (as characterized by their significant co-occurrence patterns).*

Discussion and Conclusion

While several informative associations for each community emerged, the associations with community 10 exemplify a crystal-clear exposure pathway quite unlike the others. This community is indicative of pharmaceuticals, likely in the form of direct use either for humans and/or livestock, passing through said biological systems and eventually collecting in sewage sludge (i.e., biosolids) (see fig. 4 & 5 below). It is important to remember that the communities identified in this study were derived solely from statistically significant instances of co-occurrence across all source and sink media categories between organic compounds. What this research demonstrates is the beneficial approach that graph machine learning models and network analyses lend to the identification of nuanced, often over-arching, exposure pathways and scenarios.

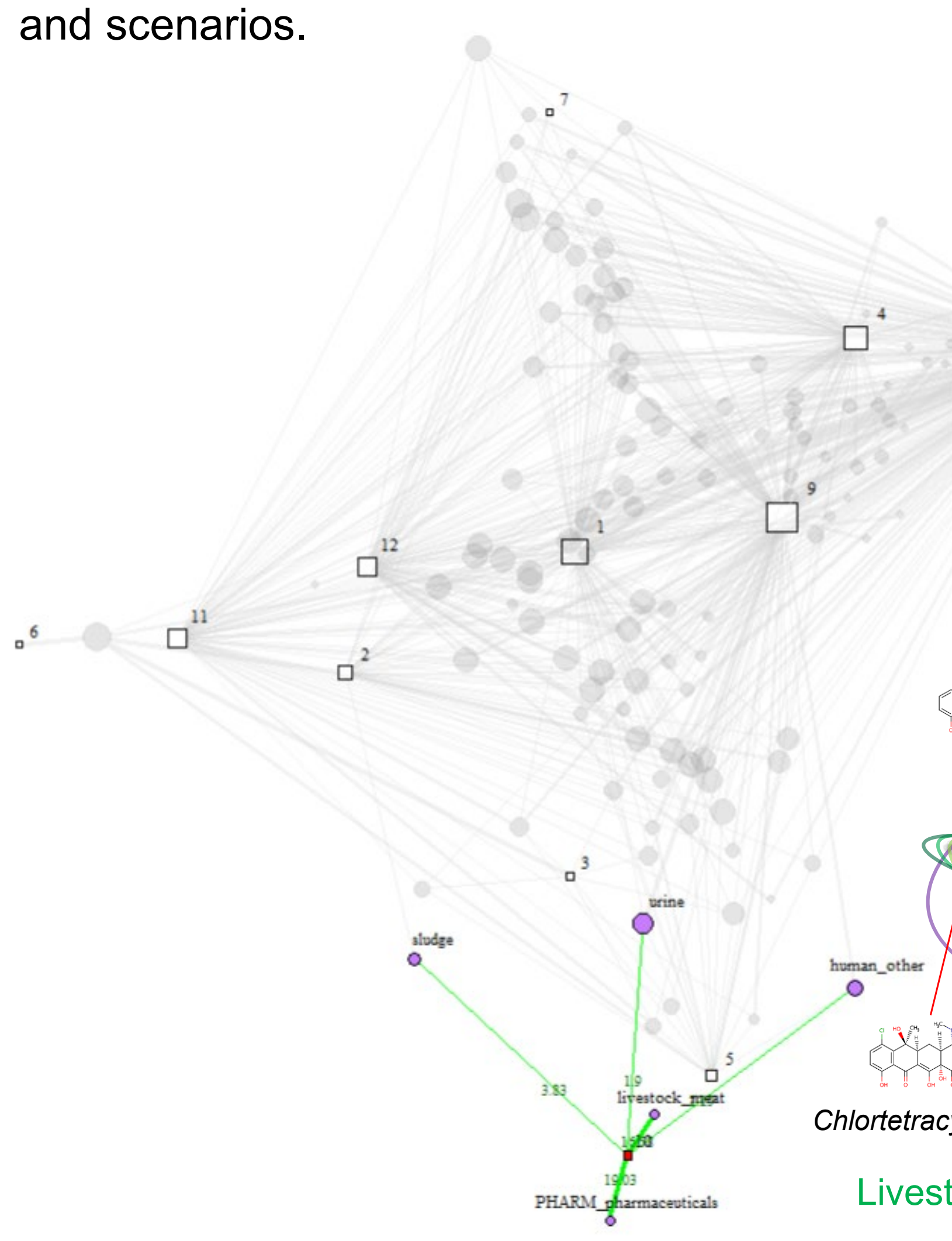
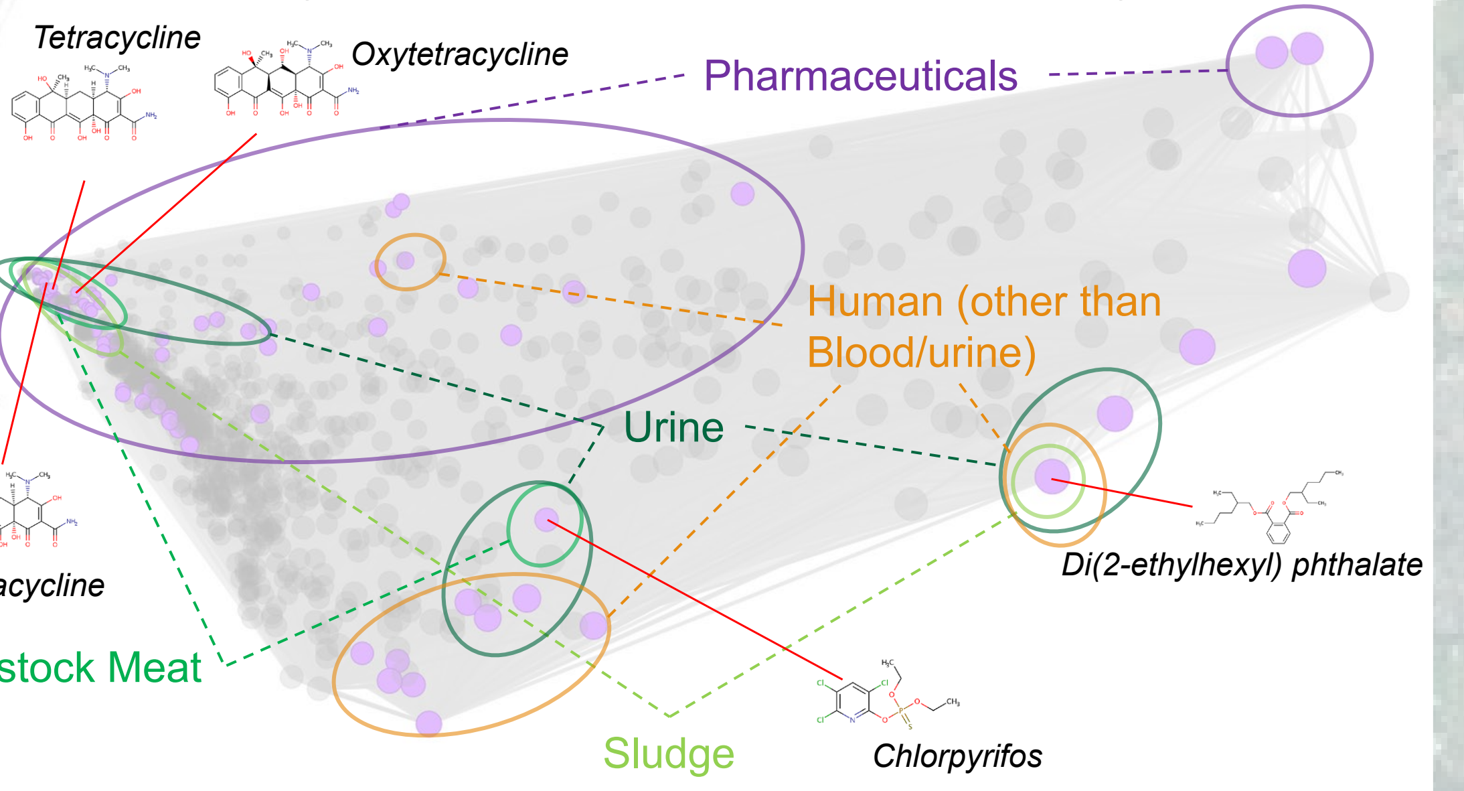


Fig 4. *(left) Correlation network of communities (squares) and media categories (circles) depicting community 10's positive coefficients as odds ratios. Node size showcases PageRank of communities and media. Node and edge color highlight community 10. Edge width is proportional to coefficent strength. Community 10 corresponds to chemicals appearing in pharmaceuticals which likely end up in human-related media (other than blood/urine), urine, sludge, and livestock meat.*

Fig 5. *(below) Original chemical-to-chemical network colored to show chemical assignment with community 10. High-degree of spatial overlap between media categories.*



Future Research

- One immediate use of this research would be source apportionment tasks in Non-Targeted Analysis (NTA)
  - Sample matrices could be cross-referenced with enriched exposure pathways
  - Once likely pathways reflecting a given matrix (e.g., "drinking water") are determined, rank ordered lists of probable confirmed-media (documented presence of suspect-chemical in media) or unconfirmed media (where no information regarding known source or sink media of suspect-chemical is present) can be generated by sorting the pathways' associated media categories by their coefficients, with the strongest, positive coefficients indicating more-probable sources

While still in an exploratory phase, the integration of this research across EPA's Center for Computational Toxicology and Exposure continues to develop. Furthermore, with the advent of continued monitoring efforts and dataset expansion and development, this analysis is designed to adapt with the expanding knowledge-base of chemical co-occurrences and, over time, will capture better-informed exposure pathways as more and more chemical presence-in-media data become available.

References:

U.S. Environmental Protection Agency
Office of Research and Development

This project was supported in part by an appointment to the Research Participation Program at the Office of Research and Development's Center for Computational Toxicology and Exposure, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.

Disclaimer: This poster does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.