Network Graph Solutions for Data Pipeline Diagnostics

Hanna Gerlovin PhD[1], William Robb MS[1], Brian R. Ferolito MS[1], Yuk-Lam Ho MPH[1],
David R Gagnon MD MPH PhD[1,2], and Kelly Cho PhD MPH[1,3]

[1]Veterans Affairs (VA) Boston Healthcare System, Boston, MA, US;
[2]Department of Biostatistics, Boston University School of Public Health, Boston, MA, US;
[3]Division of Aging, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, US

Introduction

- A **challenge** for studies using electronic health records (EHR) is the **development** of a data pipeline that **maps the raw EHR data to an analytic dataset**.
- This **data pipeline** is often a **complex network of programs**, owing to a large number of **phenotypes and data domains** that may be integrated in a single analytic dataset.
- Team members must develop and **share, on an ongoing basis, a clear understanding** of a given pipeline in order to make study progress, which can be challenging due to **varying familiarities** with existing database and process **modeling schemes**, such as Universal Modeling Language (UML) and Entity Relationship (ER) diagrams [1,2].
- ER diagrams would allow for **tracking of table dependencies**, but **not process dependencies**, which also affect **data integrity.**
- Within the **computing enclaves** of highly sensitive information, **restrictive security systems** may not allow for installation of **proprietary software** [3], limiting the availability of generalizable tools like those based on UML.
- Here we propose the use of **directed bipartite graph** models, **visualized interactively**, to address these issues and ultimately facilitate the application of advanced analytic methods that rely on data extracted from EHR databases.

Methods

- Our **novel application** of **directed bipartite graph models** allows investigators to
  - **Visualize** the **entire** data curation process
  - **Track dependency** based relevant program and data updates required when making changes to various pipeline components
  - Identify opportunities for optimization or automation of pipeline architecture
- Application:
  - A data pipeline that **links** Veterans Healthcare Administration (VA) EHR data with Medicare and Medicaid claims (CMS) data and **assembles** an analytic dataset comprised of **longitudinal patient trajectories**, including time-varying treatments, confounders, and outcomes.
  - The EHR and claims data reside on **different servers** and in different data formats, within the secure VA computing enclaves, thus both **SAS** and **SQL programming** languages are employed.
- We **implement** the graph model of the codebase using R and packages visNetwork, igraph, and shiny [4, 5, 6]. This provides an interactive visualization that:
  - Organizes a wide range of pipeline metadata
  - Facilitates exploration of the pipeline codebase
  - Surfaces diagnostics and dependency-base triggers

Evaluation of Results

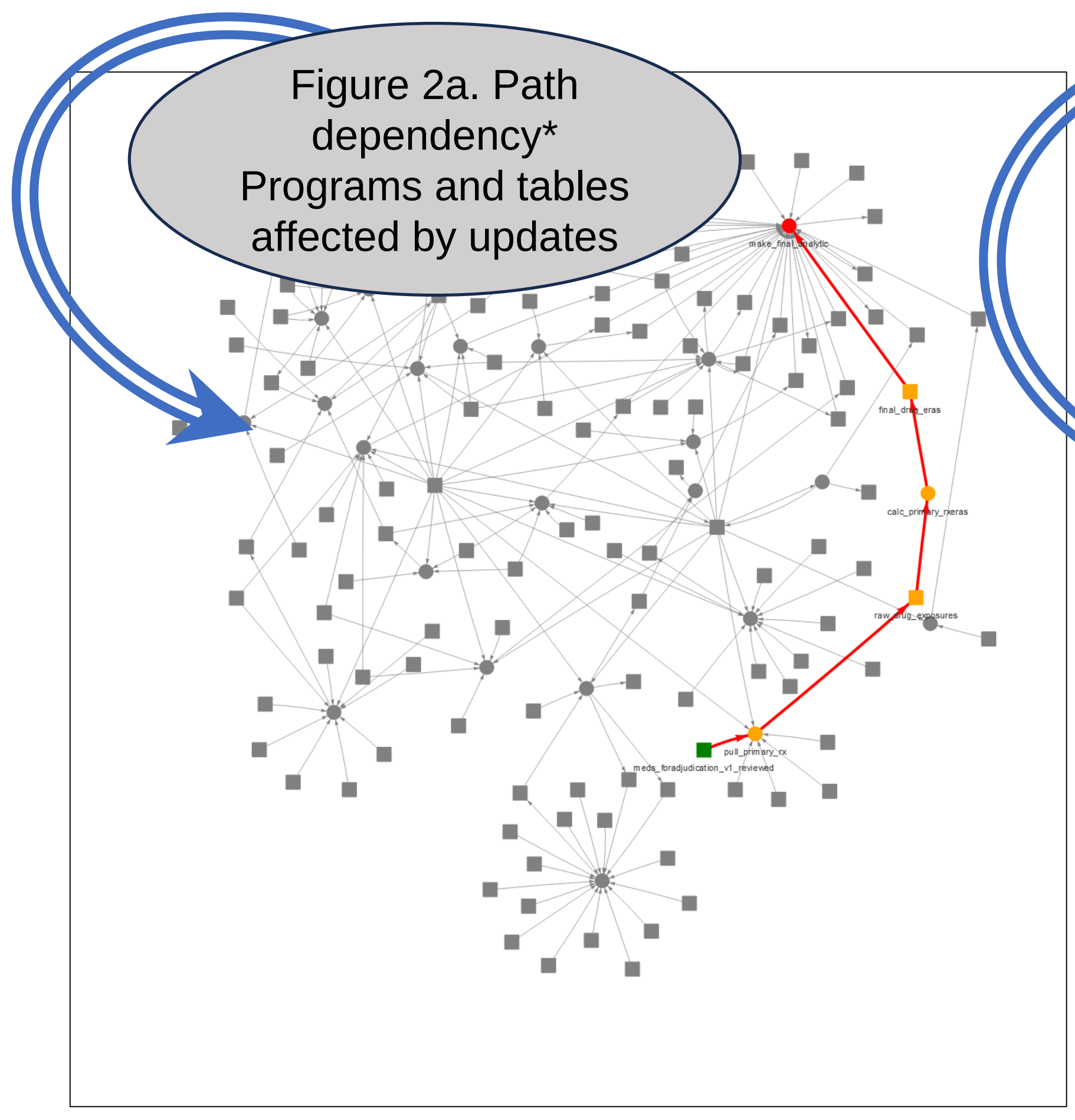


Figure 2a. Path dependency* Programs and tables affected by updates

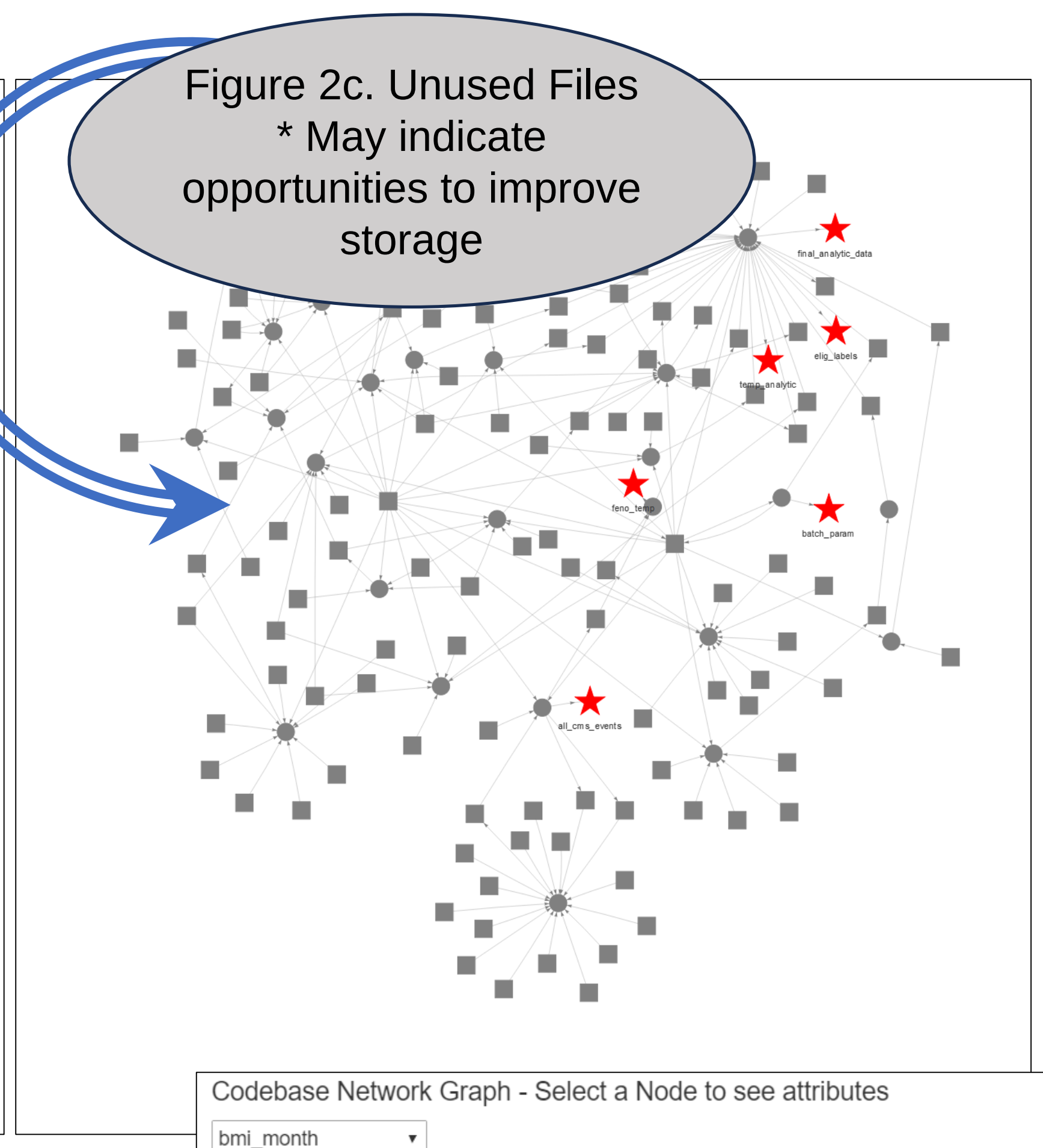


Figure 2c. Unused Files
* May indicate opportunities to improve storage

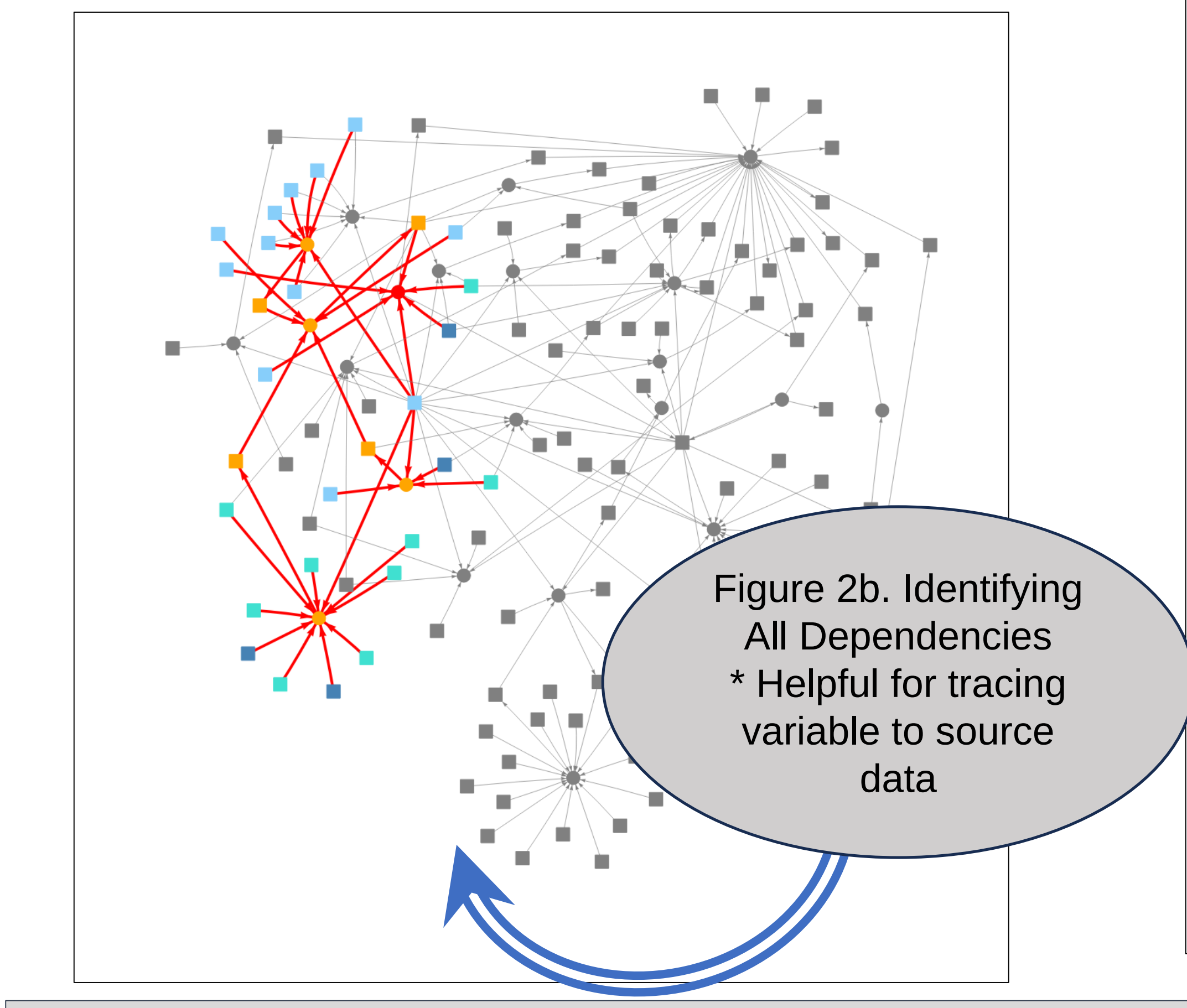
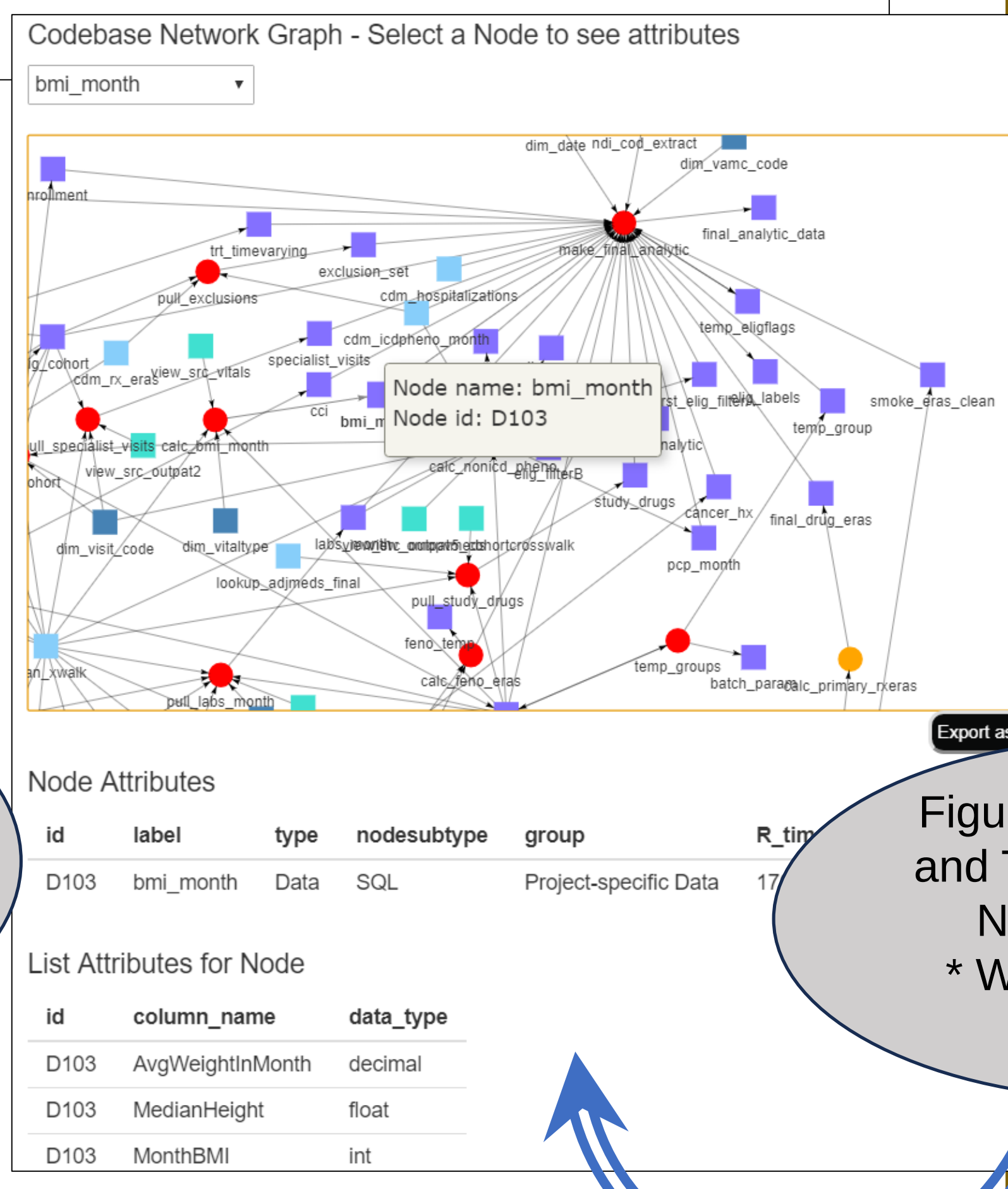


Figure 2b. Identifying All Dependencies
* Helpful for tracing variable to source data



Codebase Network Graph - Select a Node to see attributes

bmi_month

Node name: bmi_month
Node id: D103

Export as png

Node Attributes

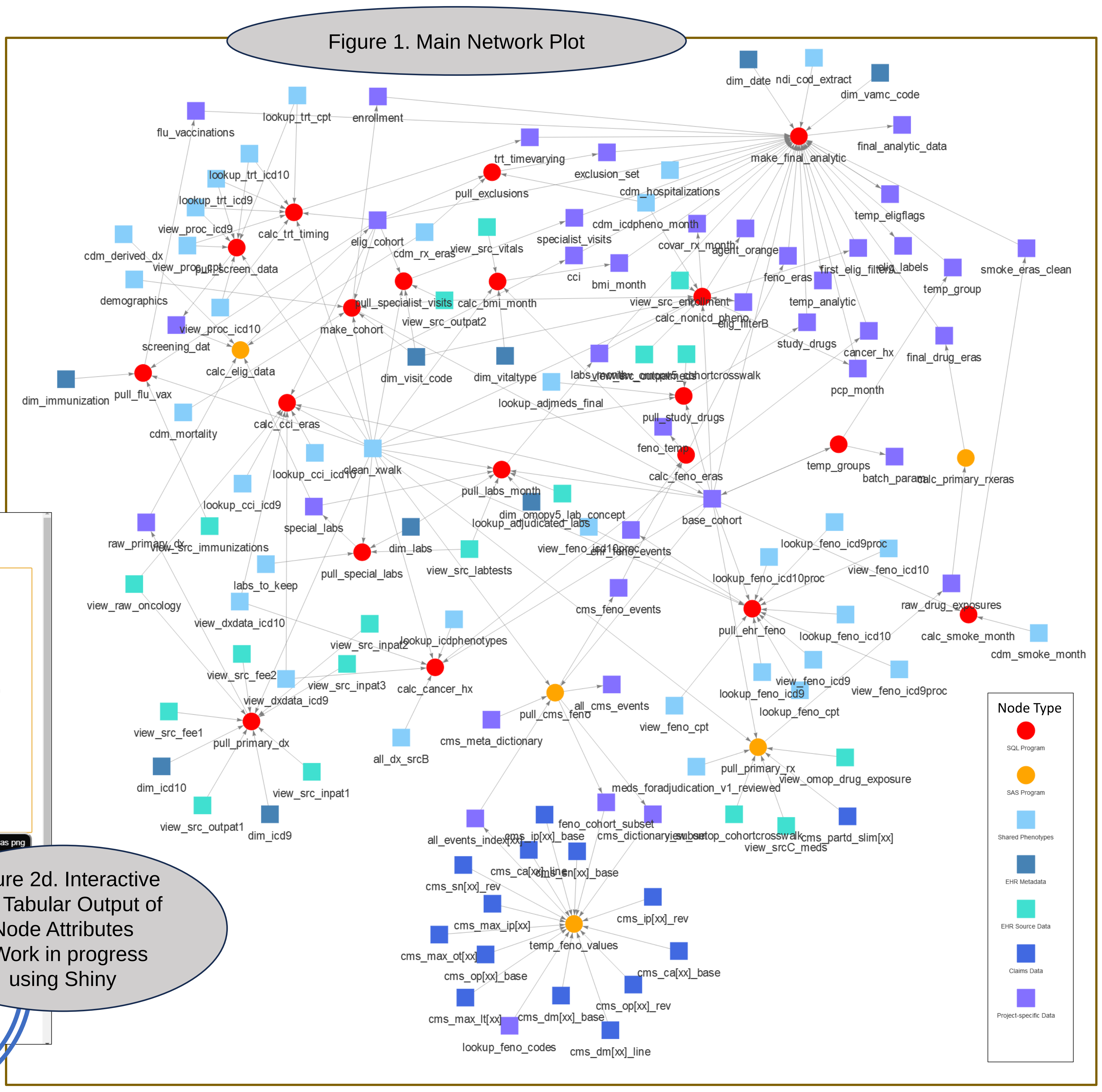
id	label	type	nodesubtype	group	R_tim
D103	bmi_month	Data	SQL	Project-specific Data	17

List Attributes for Node

id	column_name	data_type
D103	AvgWeightInMonth	decimal
D103	MedianHeight	float
D103	MonthBMI	int

Figure 2d. Interactive and Tabular Output of Node Attributes
* Work in progress using Shiny

Figure 1. Main Network Plot



Conclusions

- The representation of a data curation codebase in a bipartite network model provides a **natural way** for users to visualize, interrogate, and refine the data engineering process.
- By promoting both **transparency and data integrity**, this diagnostic approach may help **to improve the reproducibility** and quality of studies that rely on EHR and administratively-collected healthcare data.
- This framework can support tool development, process standardization, and graph-theoretical optimizations. **Future directions** of functionality include:
  - Variable tracing, incorporating semantic analysis of source programs
  - Robust quality control and verification checks, leveraging graph-theory dependency analysis
  - Flexible inspection and selection, through the continued development of an interactive interface

References

1 City, A.T.M., UML Distilled Second Edition A Brief Guide to the Standard Object Modelling Language.

2 Batra, D., Hoffler, J.A. and Bostrom, R.P., 1990. Comparing representations with relational and EER models. *Communications of the ACM*, *33*(2), pp.126-139.

3 Mitchell, E., Berkani, N., Bellatreche, L., Ordonez, C. (2023). FLOWER: Viewing Data Flow in ER Diagrams. In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2023. Lecture Notes in Computer Science, vol 14148. Springer, Cham. https://doi.org/10.1007/978-3-031-39831-5_32

4 Almend B. V. and Contributors, Thieurmel B (2002). visNetwork: A Network Visualization using 'vis.js' Library. R Package version 2.1.2, https://CRAN.R-project.org/package=visNetwork

5 Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.sf.net

6 Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2024). *shiny: Web Application Framework for R*. R package version 1.8.0.9000, https://github.com/rstudio/shiny, https://shiny.posit.co/.

Acknowledgements: HG & WR are supported by the VA Office of Research and Development Cooperative Studies VA-CAUSAL Program (CSP#2032).

Disclosures: The contents of this poster do not represent the views of the U.S. Department of Veterans Affairs or the U.S. Government. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any U.S. Government entity.