SLNET: A Redistributable Corpus of 3rd-party Simulink Models

Sohil Lal Shrestha
University of Texas at Arlington
Arlington, Texas, USA

Shafiqul Azam Chowdhury
University of Texas at Arlington
Arlington, Texas, USA

Christoph Csallner
University of Texas at Arlington
Arlington, Texas, USA

ABSTRACT
MATLAB/Simulink is widely used for model-based design. Engineers create Simulink models and compile them to embedded code, often to control safety-critical cyber-physical systems in automotive, aerospace, and healthcare applications. Despite Simulink’s importance, there are few large-scale empirical Simulink studies, perhaps because there is no large readily available corpus of third-party open-source Simulink models. To enable empirical Simulink studies, this paper introduces SLNET, the largest corpus of freely available third-party Simulink models. SLNET has several advantages over earlier collections. Specifically, SLNET is 8 times larger than the largest previous corpus of Simulink models, includes fine-grained metadata, is constructed automatically, is self-contained, and allows redistribution. SLNET is available under permissive open-source licenses and contains its collection and analysis tools.

CCS CONCEPTS
• Software and its engineering → Software libraries and repositories; Model-driven software engineering; • Computer systems organization → Embedded and cyber-physical systems.

KEYWORDS
Simulink, mining software repositories, open-source

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors(s).

ACM Reference Format:

1 INTRODUCTION
Currently there is no collection of Simulink models that is commonly used in empirical studies. Though there have been previous model collections, they lack fine-grained meta-information, are not self-contained, and are not redistributable due to restrictive or missing licenses—making them hard or impossible to use for most empirical researchers. Given the lack of such a collection, the few existing empirical studies of Simulink models have been limited to proprietary models or a small number of public models [9, 41, 42].

Deepening our understanding of Simulink models and modeling practices is important, as Simulink is a de-facto standard tool in several safety-critical industries such as automotive, aerospace, healthcare, and industrial automation—for system modeling and analysis, compiling models to code, and deploying code to embedded hardware [44, 46]. Having a large corpus of third-party Simulink models may make it easier for engineers and researchers to produce, reproduce, and validate empirical results about Simulink models, modeling practices, and tools that operate on such models.

The most closely related previous work has studied an initial collection of 391 third-party Simulink models [9] and later extended it to a curated corpus (‘SLC0’) of some 1k third-party Simulink models [11]. Boll et al. [4] collected an updated version of SLC0 and assessed the corpus’s suitability for empirical research. While pioneering larger studies and validating that models from such a corpus can be similar to industrial models, these collections consisted of a list of URLs to non-permanent resources [9] and contained models with unclear license information [11]. These collections were largely manual, which lead to inconsistencies (empty projects, duplicate projects, and missing metadata), relatively modest collection size, and may yield unintended human errors and bias.

To address these limitations, SLNET automates corpus construction and analysis, including data acquisition, cleaning (except for the rarely required manual review of a new license type), metric computation, and packaging. SLNET thereby automatically mines and analyses Simulink models from the two most popular repositories for sharing Simulink models, yielding a collection of thousands of models that is fully self-contained and allows redistribution.

To allow fine-grained selection of Simulink models and projects, SLNET computes several project-level and model-level metrics [4] and exposes them in a SQL database. SLNET similarly identifies and labels libraries and models that are test harnesses [31]. To summarize, this paper makes the following major contributions.

• SLNET is 8 times larger than the prior largest known corpus of third-party Simulink models. SLNET also adds fine-grained metrics, being self-contained and redistributable.
• SLNET [38] and its tools [36, 37] are available under permissive open-source licenses (CC BY and BSD 3-clause).

2 BACKGROUND ON SIMULINK
Simulink [23] is a widely used commercial tool-chain for model-based design [44, 46]1. Engineers typically design a cyber-physical system (CPS) model in Simulink’s graphical modeling environment. A Simulink model such as Figure 1 is a block diagram, where each block represents equations or modeling components. Depending on the block type, each block can accept input (via input ports), perform some operation on its inputs, and produce output (via output ports), which then can optionally be forwarded to other blocks via explicit or implicit connection lines (aka signal lines).

Searching for ‘Simulink’ jobs on LinkedIn in the US currently yields over 5k job postings: https://www.linkedin.com/jobs/search/?keywords=“simulink”&location-US
Simulink users can add blocks from various built-in libraries and toolboxes [25], and can also define custom blocks in "native" code (e.g., in C) using the S-function interface.

Figure 1: Sample SLNET Simulink model of a 1.5MW wind generation plant [30] with 18 blocks and 23 connections.

To deal with model size, users can create hierarchical models, by (recursively) grouping blocks in (a) a Subsystem or (b) in a separate model via Model Reference. Simulink does not permit a cyclic model hierarchy, but there may be block connection (aka data dependence) cycles, including algebraic loops.

As a first step, compiling translates the model into a toolchain-internal representation. When simulating the compiled model, the toolchain computes the output of each block at successive time steps over a specified time range using pre-configured numerical solvers. Fixed-step solvers solve the model at fixed time intervals whereas variable-step solvers automatically adjust the time intervals at which the model is solved. Simulink may reject a model if it cannot numerically solve an algebraic loop. Simulink offers different simulation modes, i.e., Normal mode "only" simulates blocks, Accelerator speeds up simulation by emitting native code, and Rapid Accelerator produces a standalone executable.

3 SLNET DESIGN & CONSTRUCTION

SLNET is not a superset of earlier Simulink corpora [4, 11] as earlier corpora were neither self-contained nor redistributable. Figure 2 gives an overview of SLNET’s construction. We built SLNET from models shared in GitHub [16] and MATLAB Central [24]. Due to time limitations we do not collect Simulink models from smaller repositories such as GitLab [17] and SourceForge [40]. Before removing projects that are empty, duplicate, or have an unclear license, a quick search for "Simulink" yields some 60 GitLab and some 70 SourceForge projects.

While GitHub offers commit-level version control, MATLAB Central “only” serves project releases. To limit SLNET’s size and due to the different versioning (git commits vs. project releases), in February 2020 we “only” collected Simulink project snapshots (i.e., all current project files plus project metadata).

GitHub provides a REST API to discover projects and extract them with their metadata. SLNET-Miner queries the GitHub API (via PyGitHub [43]) with the keyword “Simulink”. Unlike previous work [9, 11], we used keyword search and not file extension search, as file extension search is typically intended to search within a given GitHub repository and using file extension search in GitHub’s search page produced many false positives.

The GitHub API expose 23 types of project-level information [15], of which SLNET retains 20. The other 3 are redundant (full project name) or API-internal (API query relevance score and node id). From the API we also obtain each project’s topics (user-created labels and tags). From the downloaded project files, we extracted the list of Simulink model files plus the project’s license.

As MATLAB Central “only” offers an RSS feed [22] for its file exchange platform, we filter the search result feed by Simulink models and then parse the feed to collect each project’s download URL plus 14 other types of project metadata. Since from the RSS feed we could not construct the download URL for all projects, we extracted 2,941 of the 3,110 available projects.

3.1 Data Cleaning & Storage: ZIP + SQLite

We remove projects without Simulink models (i.e., file extensions slx or mdl) and projects we know to contain synthetic models (i.e., model generators [9, 10]). We heuristically search for other model generators (via terms “automat”, “random”, “fuzz”, and “generate”) in project titles, project descriptions, and project tags, which yielded 530 projects (e.g., on fuzzy logic). As we did not find evidence that these projects generate models we kept them in SLNET.

Table 1: Data cleaning: Real = has 1+ models (likely non-synthetic); License = has a license; SLNET+D = license allows redistribution; SLNET = has a model with 1+ blocks after removing potential duplicate projects; Model counts here include 1,130 library and 9 test harness models.

<table>
<thead>
<tr>
<th>Projects</th>
<th>Real</th>
<th>License</th>
<th>SLNET+D</th>
<th>SLNET</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>GitHub</td>
<td>1,284</td>
<td>232</td>
<td>231</td>
<td>225</td>
<td>2,088</td>
</tr>
<tr>
<td>MATLAB C1</td>
<td>2,941</td>
<td>2,746</td>
<td>2,728</td>
<td>2,612</td>
<td>7,029</td>
</tr>
<tr>
<td>Total</td>
<td>4,225</td>
<td>2,978</td>
<td>2,959</td>
<td>2,837</td>
<td>9,117</td>
</tr>
</tbody>
</table>
We then remove projects without a license or whose license does not allow redistribution. GitHub has a structured way for authors to set a license, which GitHub converts to a file (and exposes via an API). We manually reviewed the remaining 50 projects’ licenses (where GitHub did not understand the author’s license or for MATLAB Central projects without a BSD license).

We heuristically remove potentially duplicate projects. We consider project A a duplicate of B if (1) A and B contain the same number of Simulink model files and (2) there is a bijective mapping between models in A and B based on our Section 3.2 model metrics (excluding compile time). If A and B are from the same data source (GitHub or MATLAB Central), we keep the first-created one in SLNET. Otherwise, we keep the one from GitHub, as it offers more fine-grained meta-data. Finally, we remove dummy projects (projects whose Simulink models all have zero blocks).

Table 1 summarizes data cleaning. After removing model generators we downloaded 4,225 projects with at least one Simulink model, of which 2,978 had a license, of which 2,959 allowed redistribution. Removing 112 potentially duplicate plus 10 dummy projects yielded 2,837 projects and their 9,117 Simulink models in SLNET.

SLNET is on Zenodo (a second archive contains the 112 duplicate projects) [38]. It contains project-level information (license type, etc.) from the source repositories and the model metrics our tools extracted. Users can thus select models and projects from SLNET via SQL queries.

3.2 Project & Model Metrics

Table 2: SLNET’s project engagement distributions are long-tailed as in other studies of open-source projects [2, 18–20].

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Med</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GitHub</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stargazers</td>
<td>0</td>
<td>128</td>
<td>3.5</td>
<td>0</td>
<td>12.1</td>
</tr>
<tr>
<td>Forks</td>
<td>0</td>
<td>122</td>
<td>2.8</td>
<td>0</td>
<td>10.7</td>
</tr>
<tr>
<td>Open Issues</td>
<td>0</td>
<td>82</td>
<td>1.2</td>
<td>0</td>
<td>6.5</td>
</tr>
<tr>
<td>MATLAB Cl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comments</td>
<td>0</td>
<td>218</td>
<td>3.5</td>
<td>1</td>
<td>12.5</td>
</tr>
<tr>
<td>Ratings</td>
<td>0</td>
<td>108</td>
<td>2.9</td>
<td>1</td>
<td>6.8</td>
</tr>
<tr>
<td>Avg. Rating</td>
<td>0</td>
<td>5</td>
<td>2.5</td>
<td>3</td>
<td>2.2</td>
</tr>
</tbody>
</table>

To get an insight into the projects’ domain and popularity we first searched the user-generated project tags (i.e., GitHub “topics” and MATLAB Central “categories”) for common domains (i.e., the Simulink project domains identified by Boll et al.[4]), yielding Electronics (983), Automotive (64), Communications (61), Robotics (52), Energy (48), Aerospace (47), Biotech (20), and Medicine (2). Table 2 shows data often used as proxies for project popularity or engagement (e.g., people who have star-ed or forked a GitHub project or provided a 1–5 star rating for a MATLAB Central project). For example, a SLNET GitHub project has on average 2.8 forks.

To extract commonly used model metrics (such as number of blocks, connections, subsystems, and linked blocks) we implemented the SLNET-Metrics tool [36] on top of Simulink’s APIs. While our Simulink installation and toolbox configuration can not compile a significant portion of SLNET models (mostly due to missing toolbox licenses), these APIs still compute metrics for these non-compiling models, except for three metrics (algebraic loops, cyclomatic complexity, and compile time).

SLNET-Metrics failed to compute metrics for 88/9,117 models (21 from GitHub, 67 from MATLAB Central). Most of these 88 were due to Simulink version issues (missing Simulink toolboxes, model name conflicts with a keyword or toolbox file name) and bugs introduced by manually-edited model files. SLNET does not include metrics for these 88 models and thus also ignores them for the above duplicate-via-bijection removal.

SLNET-Metrics collects each model’s hierarchical depth, solver type, simulation mode, target hardware, and use of S-functions and model references. While SLNET models contain elements from the state-machine toolbox Stateflow, Stateflow is out of scope and our metrics do not count the Stateflow-contents of a Simulink block.

Unlike SLC0, SLNET-Metrics does not count nested blocks imported from libraries or their connections (aka “masked subsystems”). This mirrors procedural code metrics, which also do not count LOC a program imports from a library. As SLC0’s counting of such imported blocks approximates the model’s overall conceptual complexity.

SQLite is widely used, free, self-contained, server-less, zero-configuration, backwards compatible, and cross-platform: https://www.sqlite.org/index.html

Table 3: SLNET’s model metrics after removing library & test harness models; M = models; Mc = models we could readily compile; Mh = hierarchical models (readily compilable and otherwise); C = non-hidden connections; C0 = via SLC0’s metric tool; Var = variable; Nor = normal; Ext = external; PIL = processor in the loop; Ac = accelerator. For 18 models the API did not indicate simulation mode or solver type. The remaining 4 models are configured for Rapid Accelerator simulation mode.

<table>
<thead>
<tr>
<th>Source</th>
<th>Models M</th>
<th>Mc</th>
<th>Hierarchical Mh</th>
<th>Mh0</th>
<th>Blocks B</th>
<th>B0</th>
<th>Connections C</th>
<th>C0</th>
<th>Solver Step</th>
<th>Fixed</th>
<th>Var</th>
<th>Nor</th>
<th>Ext</th>
<th>PIL</th>
<th>Ac</th>
</tr>
</thead>
<tbody>
<tr>
<td>GitHub</td>
<td>1,639</td>
<td>541</td>
<td>878</td>
<td>1,304</td>
<td>190,321</td>
<td>414,241</td>
<td>188,285</td>
<td>395,725</td>
<td>860</td>
<td>762</td>
<td>1,501</td>
<td>103</td>
<td>2</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>MATLAB CI</td>
<td>6,251</td>
<td>3,636</td>
<td>3,893</td>
<td>5,566</td>
<td>838,956</td>
<td>3,197,221</td>
<td>915,975</td>
<td>3,084,605</td>
<td>1,757</td>
<td>4,493</td>
<td>5,984</td>
<td>186</td>
<td>2</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7,890</td>
<td>4,177</td>
<td>4,771</td>
<td>6,870</td>
<td>1,029,277</td>
<td>3,611,462</td>
<td>1,104,260</td>
<td>3,480,330</td>
<td>2,617</td>
<td>5,255</td>
<td>7,485</td>
<td>289</td>
<td>4</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

Understanding modeling practices would enable researchers to tune their tools to how engineers use Simulink in various settings. For example, SL Forge guides its random model generation by how often blocks appear in 391 open-source models [9]. The larger size of SLNET could thus, e.g., yield useful insights for tool design.

There may also be interesting correlations between metrics, maybe connecting model metrics to project metrics (e.g., model size metrics with project engagement). More generally, SLNET could contribute to a deeper understanding of model modularity, comprehension, quality, and maintainability [3, 12, 28, 34].

While SLNET is unlikely to exactly represent closed-source development, the precise shape of this relation is an open question. For example, for the related domain of Object Constraint Language (OCL) expressions [7], Mengerink et al. found the distribution of expression complexity mined from GitHub projects reflects the distribution in closed-source projects, so open-source projects can be used as a proxy for industrial projects [26, 27].

5 THREATS TO VALIDITY

Due to its search heuristics SLNET-Miner may miss Simulink models (e.g., by missing some of the non-documented RSS feed URLs). Furthermore, since SLNET contains only redistributable projects, results may not be representative of all open source Simulink projects. On the flip side, while removing forks and duplicates, SLNET likely contains clones (from near-duplicate projects to adapted model portions), which can be an opportunity for clone-based research (and a challenge for others). Finally, SLNET-Metrics calls the Check API of Simulink R2019b. While this API has been available since Simulink R2017b, its behavior may change across releases and thus yield different metric values in future Simulink versions.

6 CONCLUSIONS

SLNET is the largest corpus of freely available third-party Simulink models. SLNET is 8 times larger than the largest previous Simulink corpus, includes fine-grained metadata, is constructed automatically, is self-contained, and allows redistribution.

ACKNOWLEDGMENTS

Christoph Csallner has a potential research conflict of interest due to a financial interest with Microsoft and The Trade Desk. A management plan has been created to preserve objectivity in research in accordance with UTA policy. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1911017 and a gift from MathWorks.