BizGraphQA: A Dataset for Image-based Inference over Graph-structured Diagrams from Business Domains

Petr Babkin
J.P. Morgan AI Research
Palo Alto, USA
petr.babkin
@jpmorgan.com

William Watson
J.P. Morgan AI Research
New York, USA
william.watson
@jpmchase.com

Zhiqiang Ma
J.P. Morgan AI Research
New York, USA
zhiqiang.ma
@jpmchase.com

Lucas Cecchi
J.P. Morgan AI Research
New York, USA
lucas.cecchi
@jpmchase.com

Natraj Raman
J.P. Morgan AI Research
London, UK
natraj.raman
@jpmorgan.com

Armineh Nourbakhsh
J.P. Morgan AI Research
New York, USA
armineh.nourbakhsh
@jpmchase.com

Sameena Shah
J.P. Morgan AI Research
New York, USA
sameena.shah
@jpmorgan.com

ABSTRACT

Graph-structured diagrams, such as enterprise ownership charts or management hierarchies, are a challenging medium for deep learning models as they not only require the capacity to model language and spatial relations but also the topology of links between entities and the varying semantics of what those links represent. Devising Question Answering models that automatically process and understand such diagrams have vast applications to many enterprise domains, and can move the state-of-the-art on multimodal document understanding to a new frontier. Curating real-world datasets to train these models can be difficult, due to scarcity and confidentiality of the documents where such diagrams are included. Recently released synthetic datasets are often prone to repetitive structures that can be memorized or tackled using heuristics. In this paper, we present a collection of 10,000 synthetic graphs that faithfully reflect properties of real graphs in four business domains, and are realistically rendered within a PDF document with varying styles and layouts[1][2]. In addition, we have generated over 130,000 question instances that target complex graphical relationships specific to each domain. We hope this challenge will encourage the development of models capable of robust reasoning about graph structured images, which are ubiquitous in numerous sectors in business and across scientific disciplines.

CCS CONCEPTS

• **Information systems** → **Information extraction**; **Question answering**; **Multimedia and multimodal retrieval**.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591875

KEYWORDS

visual question answering, visually rich documents, deep learning dataset

ACM Reference Format:
Petr Babkin, William Watson, Zhiqiang Ma, Lucas Cecchi, Natraj Raman, Armineh Nourbakhsh, and Sameena Shah. 2023. BizGraphQA: A Dataset for Image-based Inference over Graph-structured Diagrams from Business Domains. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591875

1 INTRODUCTION

The task of visual question answering (VQA) has attracted attention in recent years due to the successes of the underlying language understanding and vision models. Open-domain VQA methods have leveraged advancements in multimodal fusion [2] and scene graphs [6] to reach near-human performance in certain sub-tasks[3]. VQA methods that are focused on human-generated input (such as documents) have taken advantage of the visual and spatial structure of such artifacts, such as the grid-like nature of forms [1] or reading-order constraints [15].

Visual graphs offer an interesting middle ground between the unique challenges of open-domain VQA, and of visually rich document understanding. They are human-generated artifacts, often used to visually convey information about various entities or concepts, and their interrelationships. As such, they do not necessarily honor the grid structure of a form, and their semantic or panoptic segmentation challenges are different from open-domain VQA.

In addition to offering unique research challenges, visual graph understanding has wide applicability in many domains, including enterprise settings. Corporate insights, supply chain data, investment networks, market interactions, and many other forms of enterprise data are often visualized as graphs. An automated approach to VQA over such graphs opens up novel opportunities for corporate entities and other users of such datasets.

In this paper, we present BizGraphQA, a VQA dataset over enterprise documents that contain graph-structured diagrams. The

[1] Sample: https://drive.google.com/file/d/1c7Ac0-tRv1-xJvArtnefMA-uMUvuaK0t
[2] Full: https://drive.google.com/file/d/1_3xs0GIuhwsT4Fg01AFyoOr0G8S_ikpr

[3] https://visualqa.org/roe.html

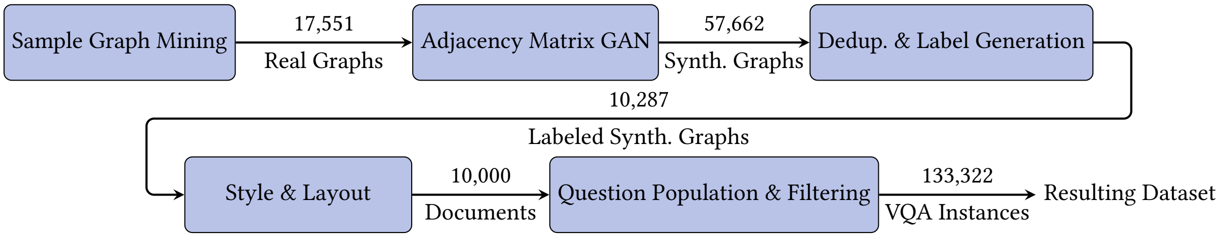


Figure 1: BizGraphQA dataset creation pipeline.

diagrams have been synthetically generated based on four domains, namely enterprise ownership charts, employee reporting hierarchies, supply chain networks, and start-up investment networks. In creating the dataset, we have followed a rigorous process of sampling, generation, and deduplication, in order to ensure diversity of visual and graphical structures, and to retain fidelity to real data without risking the exposure of confidential information.

For each diagram, we have also curated one or more questions that require spatial, visual, and semantic processing. We specifically target difficult questions that are not simply extractive and require reasoning about graph structure. Our questions cover a wide range of reasoning tasks from spatial and numeric reasoning, to single and multi-span counting. Through baseline experiments, we demonstrate that state of the art unimodal or multi-modal models struggle to perform the reasoning required by BizGraphQA questions, and fall far behind human performance. This motivates research on new VQA models that are more robust to the complexities of diverse visual reasoning tasks over graphic structures.

2 RELATED WORK

Visual Question Answering (VQA) datasets span a wide range of genres of images and corresponding question/answer pairs. The Visual QA dataset [3], composed of 200K natural images from the MS COCO collection [17], as well as 50K abstract scenes created using clipart, has been extensively used as a benchmark [12, 16, 26]. Its images are annotated with a total of over a million open-ended questions that require visual, spatial, common sense and quantitative reasoning to solve.

GQA is another dataset for question answering over natural images, that focuses on compositional reasoning over the relational hierarchy between objects in the scene [5]. Unlike VQA, GQA images are paired with scene graphs from the Visual Genome dataset [14], which are used to generate questions based on varied linguistic patterns, and to detail formal reasoning steps and annotated textual and visual cues necessary to arrive to the correct answer.

While neural models have demonstrated the ability to correctly answer questions about natural images might, numerous studies have uncovered biases that hint at the models exploiting statistical regularities in the question/answer distributions or picking up on universal cues in the images themselves, rather than performing any actual visual reasoning over the image's scene [5, 7, 18]. Following that motivation, several synthetic datasets have been proposed to allow for the full control over the answer distribution, and the scene — as to make questions truly challenging for existing VQA systems. One example is CLEVR [7], which is comprised of 3D renders of solid shapes of different materials positioned in various spatial configurations in the vein of Terry Winograd's classic system SHRDLU [28]. The authors' main focus is on evaluating complex reasoning in a controlled environment isolated from exploitable biases and heuristic cues. IconQA [18] consists of abstract images similar to those in VQA, sourced from early education textbooks. Scenes and questions in this dataset are designed to be unsolvable via mere object recognition and rely on common sense reasoning.

An important branch of VQA is that of DocQA where questions are grounded in visually rich documents. DVQA [8] and FigureQA [9] focus on questions about data visualizations such as line plots, pie charts and bar charts. The aim of these datasets is to probe the model's ability to learn relationships between underlying quantities and their varying graphical representations. InfographicVQA [19] takes this a step further by focusing on information dense infographics that combine graphics with text, tables and data visualizations. They source about 5K digital-born infographics from the Internet and have them crowd-annotated with about 30K questions following a methodology similar to SQuAD [24]. Question types encompass multi-span image-span and non-extractive answers that must be gleaned from multimodal sources. The authors demonstrate that traditional VQA performs poorly in this high information density and also challenging layout setting. However, the fact that baseline systems such as LayoutLM [29] manage to achieve strong performance despite lacking the ability to explicitly model graphical structures suggests the datasets are prone to heuristics.

This prompts us to look for difficult datasets that is hard for both VQA and DocQA systems.

Similar in spirit is the AI2D dataset [10] which comprises 25k graph diagrams from about 20 different domains from middle school science textbooks. In addition to images, the dataset provides structured annotations of each diagram - nodes, texts, arrows, as well as bounding boxes. Answering these questions requires the ability to reason about the structure of a diagram and the semantics of its constituents and their relationships. Their baseline approach is a complicated pipeline that first parses diagrams into entities and relationships then does some attention passes over the content of different elements of the diagram, effectively rendering the task as that of multiple choice NLI.

The dataset that comes closets to our work is FlowchartQA [27] with comprises almost a million algorithmically generated graphs. FlowchartQA questions are harder by design and require reasoning about graphical relations. However, FlowchartQA questions are generic and are posed over abstract synthetic graphs – with no relation to a real domain.

3 DATASET DESCRIPTION

BizGraphQA is composed of graph-structured diagrams based on four domains: Company Organization Diagrams, Employee Reporting Hierarchies, Supply Chains Networks, and Startup Investment Networks. In addition to business significance, these domains have unique characteristics that make them suitable candidates for a complex VQA task: 1) each domain embodies sufficiently complex relations among entities, 2) the semantics of those relations are distinct from domain to domain, and 3) each domain exhibits interesting structural patterns that are unique to it. In the following sections we will describe each domain in further detail and explain our synthetic sample generation approach outlined in Figure 1. For each domain, we begin with a collection of real-world data, which we describe in minimal detail in order to maintain anonymity and confidentiality. Next, we will describe how the real-world data was used to generate the synthetic samples so as to maintain high fidelity to the complexities of the original dataset.

3.1 Overview of Business Domains

Organization Structure **(ORG)** — We curated a collection of organization hierarchy diagrams spanning 8 enterprise domains including corporations, financial institutions, and government agencies. While corporate clients produce this information in the form of non-standard image documents, internal knowledge workers convert them into machine readable diagrams, exportable in json format. The original images vary in quality and structural complexity — often spanning over multiple pages with call-outs to different parts of the diagram. By contrast, the converted graphs are clean and considerably simpler in terms of structure since usually only a subset of the complete organization structure is relevant for client due diligence purposes. Nonetheless, these machine readable graphs represent interesting structures, which we specifically resample for, as described in the next section. The graph represents, for a given company, a complete structure of business owners and subsidiaries, with the breakdown of share of ownership among them. Consequently, the nodes of the graph represent legal entities while edges stand for directional ownership with an associated degree of ownership that can be either explicitly indicated in the diagram, contained in the footnote, or implied (such as whole-ownership). The structure of these diagrams usually forms a tree-like hierarchy, where the entity at the root constitutes an ultimate parent. However all entities in the hierarchy need not have a common ancestor (see Figure 6). As part of client due diligence, auditors might seek to answer certain questions using this diagram, e.g. to determine the ultimate parent of a given entity in the hierarchy, along with its intermediary owners and the overall degree of ownership. Answering these basic question enables them to perform business-critical checks on the client, for example, whether the given company is owned by a publicly traded firm, a foreign individual or a government entity.

Enterprise Management Hierarchy **(MGMT)** — We sample a subgraph from the internal hierarchy of reporting lines, encompassing employees at all levels of the organization. Nodes contain person names along with their title. Edges can specify either a line manager or a local manager (multiples allowed). Typical questions include finding out a given employee's manager or the number of employees in a certain organization, team, or business unit.

Supplier Networks **(SUPP)** — Similarly, we collect supply chain network data from a proprietary database. Each node represents a company and directed edges represent supplier/customer relations. Edges can indicate the investment volume or share. Supply chain network data is commonly used to perform risk analysis based on geopolitical disruptions, opportunity analysis based on multi-hop relationships, or governance analysis based on local or regional regulations.

Startup Investment Activity **(INVST)**— Lastly, we collect data on venture capital, private equity and M&A activity from a proprietary database. We focus on M&A activity where nodes are either a startup or an investor and edges are events such as funding rounds or acquisitions, accompanied with a date and, if applicable, a funding amount. This collection opens up the possibility of asking a variety of interesting questions, including weak and strong investment relationships, amounts raised in each round, the investment behaviors of major firms, and the fundraising journey of various sectors.

3.2 Mining Challenging Graphs

The first step in generating a visually and semantically challenging dataset is to ensure that each of the four collections include a sufficient number of diverse graphs. The ORG dataset comes with about 26K individual graphs, but for each of the remaining datasets, we aimed at sampling a collection of about 9,000 viable subgraphs from one large graph structure.

The MGMT graph is composed of about 300K nodes. We sample from this graph by selecting a random node then traversing to the maximum depth of 3 nodes in both the incoming and outgoing directions, using the utility functions provided by the Python `networkx` library[4], namely `bfs_predecessors` and `bfs_successors`.

For INVST we started with about 120K nodes representing either startups or investors, connected via 400K edges representing funding or M&A events. Similarly to our approach for the MGMT graphs, we picked a random node (which could be either a startup or an investor) then traversed to the max depth of 2 in the descendant direction and 1 in the ancestor direction. The rationale for sampling shallower graphs was to limit the growth in the number of nodes, which in this domain we found are much more densely connected. We observed that startup-investor interactions tend to be wide and shallow, e.g. many investors fund few startups but are otherwise not connected via a common ancestor.

For the final domain of SUPP, we began with a master graph of 455K relations. We sampled subgraphs from it using the same procedure with max depth of 3 in each direction and filtered out graphs with less than 5 nodes. We found the following check greatly improved the quality of resulting graphs by promoting symmetry: the degree of highest degree node has to be between 4 and 10, and the difference between incoming and outgoing edges should not exceed 2. For all domains, we clipped the maximum number of nodes to 28.

To ensure sufficient diversity and complexity of graphs to train our generative model, we analyzed the properties of the sourced

[4] https://networkx.org

graphs. The initial set of ORG graphs was heavily skewed towards trivial linear graphs with 5 nodes or less. In order to promote structural diversity in our training data, we first filtered out graphs with fewer than 3 nodes and those having a maximum degree of less than 3. Next, we applied random oversampling based on the graph's maximum node degree to ensure the model's even exposure to graphs of different complexity.

For the remaining three domains, given that we had greater control over the sampling procedure and, as a consequence, were able to acquire higher diversity graphs, no additional rebalancing was necessary. The only post-filtering was applied to the MGMT dataset, to discard slightly over-represented linear structures.

3.3 Adjacency Matrix Synthesis using GAN

In order to generate synthetic graphs with structures that faithfully reflect that of the target domain, we employ the standard DCGAN image generation model [23] implemented in tensorflow2[5]. We represent graphs as images by converting them into an adjacency matrix, post-padded to the maximum size of 28x28. We fixed the size of the noise dimension at 100, batch size at 256, and the number of training epochs at 500.

The generator's output is a single-channel 2-D image of `tanh`-activated values. Accordingly, we converted the training adjacency values into the [-1, 1] range and we found it helpful to additionally apply uniform label smoothing with the mixing factor of 0.5. At training time we applied a dropout of 0.3 to the discriminator and used Adam optimizers [13] for each network with the learning rates set to 1e-4. During training, in addition to the losses, we monitored the visual appearance of synthetic samples compared to real ones, resuming the training for an additional 500 epochs if necessary. We did not perform any hyper-parameter tuning.

After the training is finished, we iteratively generate batches of synthetic samples up to a target number of examples. We binarized predicted adjacency matrices by thresholding the output values at 0 and used them to initialize `networkx.DiGraph` objects. We also used the discriminator's output as a criterion for selecting realistic graphs, using a threshold chosen empirically for each domain: 0.5, 0.25, 0.5, 0.6 for ORG, MGMT, INVST, SUPP, respectively. Subsequently, we applied postprocessing to the resulting graphs, some of which were common, whereas others were domain-specific:

1. Remove zero-degree nodes and nodes with self-loops (all domains).
2. Select the largest connected component in the graph (all domains).
3. Remove nodes causing cycles if the domain does not allow them (all domains except for SUPP).
4. Force common ancestor by identifying a common root via topological sort and selecting all nodes reachable from it via breadth-first search (MGMT only).

The resulting graphs were saved in `pydot` format[6]. Lastly, to ensure there were no redundant graph structures, we used the `networkx` function `is_isomorphic` to identify and de-duplicate identical graphs. Basic statistics of the resulting graph population are summarized in Table 1.

	ORG	MGMT	INVST	SUPP
Number of graphs	4,100	1,500	1,500	2,900
Median node count	10	10	11	13
Median edge count	10	9	12	19
Median mean degree	2	1.8	2.15	2.9
Median max degree	4	4	7	8

Table 1: A breakdown of generated graphs by domain.

3.4 Synthetic Node and Edge Labels

ORG. The graphs in this collection require nodes with realistic company names, and edge labels that indicate ownership breakdowns. To generate the node labels, we tokenized the company names in the real dataset and created random juxtapositions of tokens to generate synthetic names. The tokens were sampled based on their relative frequencies in the real dataset. It is not uncommon for companies in the same hierarchy to have similar names. Therefore we allowed a 25% chance that all companies in the diagrams will share a common prefix. Next, we appended random business entity types using a predetermined set (e.g. "LLC", "Inc.", etc.), as well as location terms using the `us` library[7]. The resulting company names tend to be fairly long, and sometimes have substantial overlap among different subsidiaries, which we believe to reflect the reality of the domain and present a challenge for VQA systems. Labeling the edges was a matter of assigning ownership breakdowns to each graph. For multiple outgoing edges we iteratively sampled a percentage so that all sum to 100%. In addition, we ensured that ownership data was represented in diverse ways. For instance, in some random diagrams we removed a subset of edge labels and added a legend containing a variation of the text "All subsidiaries are 100% owned unless otherwise noted".

Lastly, for each graph we generated a title consisting of one of a few generic templates populated with the root company's name[8].

MGMT. The nodes in a management hierarchy diagram are labeled by names and professional titles. Using the `names` library[9], we generated random names, concatenated with random titles. To make labels consistent, titles were generated in accordance with their distributions in the original dataset, such that title of child nodes are sampled from a distribution conditioned on the parent's title. This way a senior manager, for example, is unlikely to report to a junior analyst. The edges can indicate main, alternative, or local managers, therefore edge labels were generated to indicate one of these exclusive categories. We omitted the main supervisor label with a 30% chance per diagram.

INVST. The nodes in this collection represent investors or startups, and the edges represent investment activities (aka "deals"). To generate node labels we again sample company names, this time without encouraging a common prefix. For each relation we sample from a range of deal dates and deal sizes, and restrict the deal type to one of the four most frequent categories: "Early Stage VC", "Buyout/LBO", "Later Stage VC", "Seed Round". Additionally, we

[5] https://github.com/hcnoh/DCGAN-tensorflow2

[6] https://github.com/pydot/pydot

[7] https://pypi.org/project/us/

[8] The root company often reflects the ultimate parent in the hierarchy.

[9] https://pypi.org/project/names/

generate a title depending on the type of the root entity (investor or startup).

SUPP. Supplier networks represent supplier/customer relationships, where each node is a business. To label the nodes we sample company names in the same way as Organization Hierarchy/Startup graphs. We label the edges according to the volume of supplier relation in USD. Such relationships often follow a long-tail distribution. So in order to mimic this we sample from a Pareto distribution, scaled by the maximum value encountered in the dataset. Lastly, we use multiple suffixes: “k”, “m”, “bn”, “trn” to denote respectively thousands, millions, billions and trillions, respectively.

3.5 Styling and Typesetting

The resulting graphs, despite having diverse structures and labels, have largely the same appearance if rendered using the default dot theme. Therefore we inject further visual diversity into each collection. For the MGMT domain we implemented a custom styling logic which randomizes node shapes and colors, as well as edge thicknesses[10]. For the remaining domains, we sourced 11 themes from the GraphViz gallery page[11] and programmatically copied their styling attributes onto our graph .dot files. This allowed us to markedly diversify the look of the synthesized graphs and to choose which types of layout worked best for each domain. While each particular domain might have a certain preference for particular node, edge, and layout styles, we chose not to differentiate among styles for any of the domains, in order to keep the dataset challenging, and to encourage the VQA models to learn style-invariant representations.

In real-world settings, the diagrams are usually embedded inside a document. This would pose an additional challenge in understanding the diagrams because a recognition system must now address complex layouts, overlapping features and potential noise introduced by the various document elements such as captions, headers and even other figures. To simulate this setting, we use a synthetic document generator [25] that automatically produces realistic documents containing the diagrams in different layout structures. Specifically, the generator uses stochastic templates that can model different layouts, diagram descriptions, section titles, tabular formats and header/footer artefacts, along with various style attributes such as font, color, spacing, etc. Some of those elements we populate from the graph metadata, such as titles and legends, while others like footers are placed randomly to promote robustness to noise and to discourage heuristics. We utilize this generator to render a broad range of layout templates enclosing these diagrams, thereby enabling a higher level of complexity induced by the diversity in visual appearance.

Figure 2 shows an example of a *document embedded diagram* and illustrates the contrast between treating the diagrams in isolation. In particular, we can observe that the border lines decorating a diagram could be mistaken for other line strokes in it or a new challenge arises in the form of associating a diagram with its title and caption.

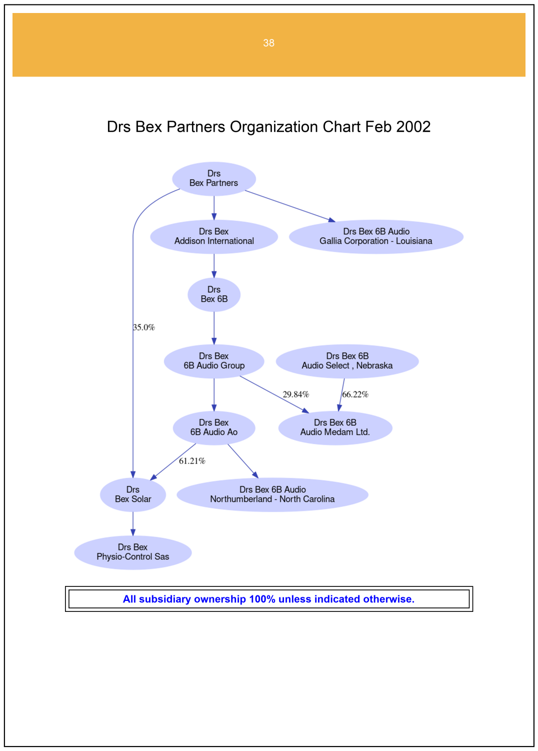

38
Drs Bex Partners Organization Chart Feb 2002
Drs Bex Partners
Drs Bex Addison International
Drs Bex 6B Audio Gallia Corporation - Louisiana
Drs Bex 6B
35.0%
Drs Bex 6B Audio Group
Drs Bex 6B Audio Select , Nebraska
29.84%
66.22%
Drs Bex 6B Audio Ao
Drs Bex 6B Audio Medam Ltd.
61.21%
Drs Bex Solar
Drs Bex 6B Audio Northumberland - North Carolina
Drs Bex Physio-Control Sas
All subsidiary ownership 100% unless indicated otherwise.

Figure 2: An example of a styled diagram typeset in a document.

3.6 Question Template Design and Population

The last step is to generate diverse and challenging questions. To generate questions, we opted for a template-based approach targeting about 15-20 templates per domain. We solicited question templates from a domain expert, instructed to provide realistic questions specific to each domain, while replacing entity names and numeric values with placeholders. Answers to these questions had one of the following types: 1) a span or multiple spans of text contained in the diagram (usually referring to entities or edge labels), 2) yes or no – probing whether certain entities in the question satisfy a particular configuration, 3) a count e.g., of the entities of a particular type.

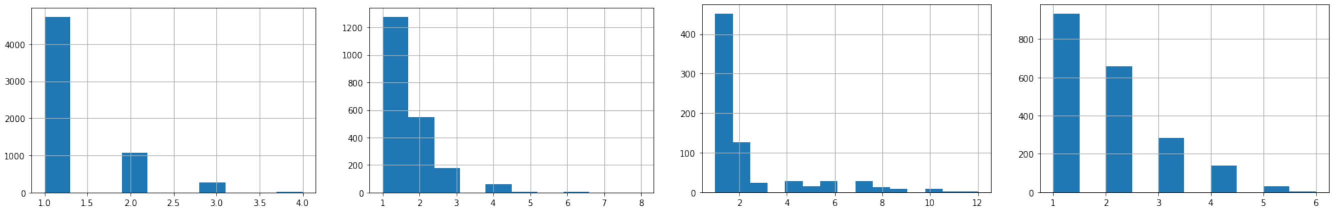
Given that we have access to the underlying graph data structures for each diagram, we were able to easily parameterize and readily obtain gold answers for each template programmatically, using networkx functions. For example, to compute the answer to

	ORG	MGMT	INVST	SUPP
Single Span	1 (100%)	2 (100%)	8 (100%)	5 (91%)
Multi-Span	2 (73%)	5 (27%)	1 (43%)	1 (72%)
Yes/No	6 (40%)	2 (58%)	5 (27%)	2 (14%)
Counting	6 (52%)	12 (57%)	3 (55%)	7 (73%)

Table 2: Counts of question templates by answer type. Percentages initially answerable indicated in parentheses.

[10] The MGMT collection is the most “tree-like” collection in our dataset, so applying a bespoke diversification algorithm allowed us to counteract its structural simplicity compared to other domains.

[11] https://graphviz.org/gallery/



(a) ORG **(b) MGMT** **(c) INVST** **(d) SUPP**

Figure 3: Histograms of the number of spans by domain.

the question "Who is the ultimate parent of Company X?" we begin by identifying the root node of the graph using topological sort, then sampling a random node among its descendants and use its label to fill in the slot in the question, and the label of the root node as the answer. After implementing approximately 70 templates across all domains and answer types, we iteratively verified a small sample of resulting instances for each template, and made necessary adjustments to the logic to ensure correctness of generated answers.

Examples of diagrams along with questions across different domains and answer types are given in Appendix A.

3.7 Subsampling of Unanswerable Instances

Given the random element in the question generation, it is possible to produce a large number of unanswerable questions, depending on the complexity of the question and the diagram. The definition of a null-answer is specific to each answer type for example, it is an empty string for single span questions and an empty list for multi-span questions. For counting questions we consider zero as the null-answer and false for yes/no questions. Table 2 provides a breakdown of implemented question templates by answer type along with the fractions of answerable instances per each type.

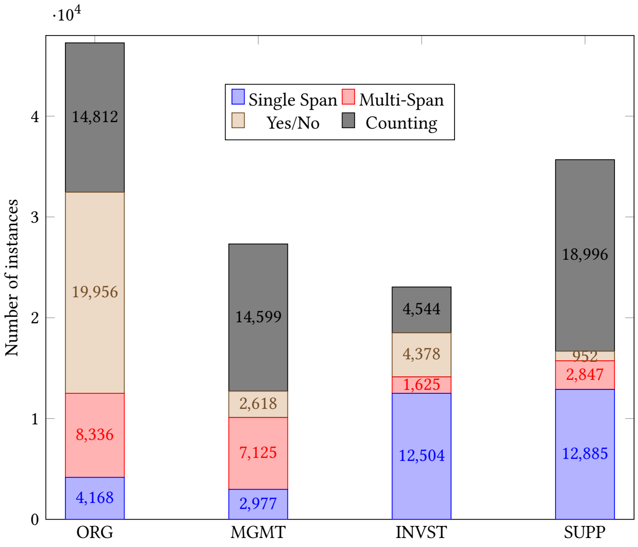


Figure 4: Final counts of question instances by type.

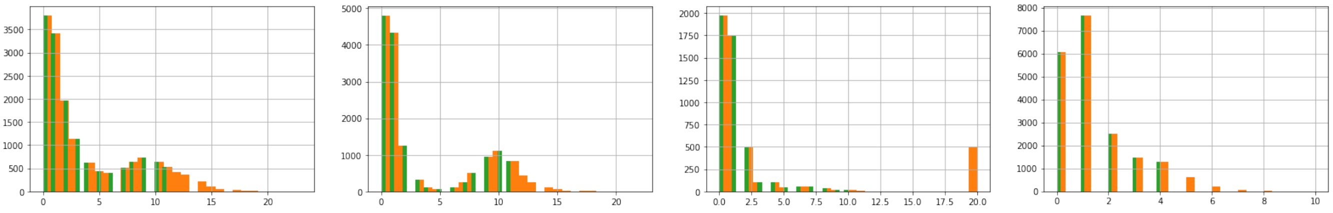
While some percentage of null-answers is acceptable, it would be undesirable for them to constitute the majority of answers for any given question type, as is especially the case for multi-span questions in MGMT and yes/no questions in INVST and SUPP. Thus, we resample question instances as follows. For single and multi-span questions we downsample questions with null-answers to not exceed the number of questions of the corresponding type having non-null answers. For binary yes/no questions we sample positive and negative answers with equal proportions. Finally, for counting questions, we downsample the 0 class to not exceed the second most common class by more than half of one standard deviation across non-zero classes. Figure 4 provides summary statistics of the resulting question population after resampling.

3.8 Analysis of Answer Distributions

To gain some insight into the complexity of questions and the diversity of answers in each domain, we examined the space of generated answers for each question type. Answers to the ORG's only single span question comprise 4168 distinct answers, all of which are company names, each occurring exactly once. Answers to MGMT's two single span questions total 1414 distinct answers, with few frequently occurring answers being generic employee titles, followed by a long tail of person's names. INVST has the largest number of single span questions, resulting in a set of 6358 distinct answers made up of few frequently occurring generic deal types followed by a long tail of company names and dates. Finally, answers to SUPP's five single span questions total 5764 distinct strings, a large portion of which are frequently repeating relationship values, similarly followed by a long tail of company names.

Additionally, for multi-span questions we provide the histrograms of the number of spans in each domain (Figure 3). In ORG and MGMT domains, the majority of multi-span answers do not exceed 3 spans, whereas in INVST and SUPP, due to more densely connected graphs, answer length can reach as many as 10 spans. Still, answers comprising just one span are prominent in all domains, accounting respectively for 78%, 61%, 62% and 46% of all multi-span answers in each domain. When combined with single span answers, in total those account for nearly 90% of all extractive questions (87%, 84%, 98%, 93%, respectively for each domain).

Next, we examined the distribution of answers to counting questions (Figure 5). The maximum value of the answer is upper-bounded by the number of nodes in the corresponding graph but it is also dependent on the specificity of the question e.g., counting



(a) ORG **(b) MGMT** **(c) INVST** **(d) SUPP**

Figure 5: Distribution of answers to counting questions by domain.

the total number of entities vs entities satisfying particular conditions, which varies by domain. We observe that, for the most part, the frequency of each answer value is inversely proportional to it magnitude. We clip the maximum value by the 90-quantile as is indicated in the diagram in green. We did not conduct any analysis of yes/no questions, since their answers were equalized, as described in Section 3.7.

4 BASELINE EVALUATION AND DISCUSSION

To demonstrate the challenging nature of our graph question answering dataset, we finetune the popular SotA model, LayoutLMv2 [30], which has historically shown strong performance on the Document VQA leaderboards[12]. LayoutLMv2 leverages a powerful visual backbone and a position-aware text encoder, both of which reasonably equip it for our multi-modal task. Evaluating purely text-based or graph neural network baselines is left out of the scope of this paper, but given the unimodal (text) or bimodal (text+spatial) nature of those models, LayoutLMv2 is expected to outperform them, as is demonstrated in Xu et al. [30].

In performing the evaluation, our goal is to gauge model performance in a realistic end-to-end setting over raw images without access to the underlying graph data structures. We opt for preprocessing the images using `Tesseract`[13] and Google Cloud OCR service[14], which is standard for LayoutLM family of models, and which allows us to emulate the practical challenge of possibly degraded model performance due to OCR errors[15].

Depending on the question type, we devised different prediction heads: extractive QA for single and multi-span predictions, as well as single logistic and multinomial softmax outputs for yes/no and counting questions, respectively, using the Huggingface implementation[16]. For each question type, we finetuned LayoutLMv2BASE along with an appropriate prediction head for 3 epochs, using a Tesla T4 GPU with a batch size of 8 and without early stopping. We perform a standard 75/25 train test split with no stratification and report results on the test set. Our goal is to measure the "best-case performance" of the model. Therefore, instead of an end-to-end scenario, we train a separate model head for each question type, and evaluate each model type separately, foregoing the upstream task of question type classification.

To evaluate single-span extractive predictions, we report exact match (EM) along with average normalized Levenshtein similarity (ANLS) [20] between the gold and predicted spans. Since we observed earlier that among nominally multi-span question instances, as much as 90% consist of only a single element, we treated them the same as single-span. Proper evaluation of true multi-span predictions is challenging since LayoutLMv2 does not have a built in capability for multi-span QA. Thus, given their relative infrequency, we chose to assign partial credit in cases of when the predicted span matched one of the gold spans – by dividing the score of the best match by the total number of gold spans. Concurrently, we explored a token classification-based alternative, which would enable multi-span predictions, but its inferior performance even on single span questions prompted us to revert to the QA model as a stronger baseline. For yes/no questions we report overall accuracy, binary precision, and recall. Finally, for counting questions, we similarly report overall accuracy but macro-aggregate precision and recall across all categorical labels. Table 3 summarizes our experiments with the baseline on each domain and each question type.

The baseline's performance on extractive questions is highly variable across domains, ranging from the exact match rate of just 14.5% for SUPP to as high as 86.8% for MGMT. We attribute high scores on MGMT to several factors. First, graphs in this domain are some of the simpler ones in terms of node and edge counts, as well as node degrees, as is reflected in Table 1. Second, the space of answers of this domain is the smallest among the four, with a large portion of answers being generic employee titles drawn from a small set (Section 3.8). The remaining answers are person's names that are more diverse than titles but usually constitute shorter spans than company names in the other three domains. Third, MGMT domain has the highest rate of null-answers, which could account for a substantial portion of the model's correct predictions. Nonetheless, upon manual examination, the predictions looked consistently good. Given this domain's relatively simple graph structures and questions predominantly targeting single-hop relations (e.g., "one's

	Extractive		**Yes/No**			**Counting**		
	EM	ANLS	Acc	Prec	Rec	Acc	Prec	Rec
ORG	51.0	67.4	49.4	49.4	100.0	22.1	1.8	8.3
MGMT	86.8	90.5	49.9	49.9	100.0	29.8	2.5	8.3
INVST	34.2	59.9	51.5	51.5	100.0	38.1	3.2	8.3
SUPP	14.5	51.4	52.9	52.9	100.0	41.1	8.2	20.0

Table 3: Baseline performance breakdown.

[12] https://rrc.cvc.uab.es/?ch=17&com=evaluation&task=1
[13] https://github.com/tesseract-ocr/tesseract
[14] https://cloud.google.com/vision/docs/ocr
[15] We observed Google Cloud OCR to produce markedly better results compared to Tesseract.
[16] https://huggingface.co/docs/transformers/model_doc/layoutlmv2

direct manager"), it would be interesting to examine the extent to which LayoutLMv2 simply relied on proximity heuristics, and also to investigate the amount of additional complexity needed to render them unreliable.

The second highest scoring domain, ORG, has a dramatically lower exact match rate of 51 and a normalized Levenshtein similarity of 67.4, despite similar graph complexity. One factor is undoubtedly the larger vocabulary of possible answers and the absence of generic repeating answers. Also, relations captured by the questions in this domain are less local than they are in MGMT (e.g.,"ultimate parent", "intermediary parents" – that could connect nodes multiple hops away from each other). Still, the low number of extractive question types for this domain likely contributed to it being easier for our baseline than INVST and SUPP.

The last two domains, INVST and SUPP, have both the largest numbers of single span questions and the largest sets of distinct answers, which is reflected in their progressively lower scores[17]. Questions in these domains are diverse and include both single hop relations (e.g., "In which round did Company A fund company B?", "What is the volume of X sourcing from Y?") as well as multi-hop relations requiring aggregation across multiple entities (e.g., "Which company received a highest single investment?", "What is Company Z's top supplier?"). In addition, questions in the SUPP domain refer to some of the most dense and complex graphs in the entire dataset.

When it comes to binary yes/no questions, the baseline shows consistently poor performance, matching the base rate of the positive class. It is disappointing that the notable differences between domains, graphs, question types and counts do not seem to affect the performance in any meaningful way. The performance is similarly low on counting questions, where the recall appears upper bounded by the number of possible values, while the overall accuracy reflects the proportion of the majority class (most often either 0 or 1). Having experimented with few hyperparameter settings, we conclude the last two tasks are not properly learned without significant modifications to the architecture, the dataset, or the training procedure. This suggests LayoutLMv2 is most suitable for extractive tasks, and may not be the best choice for fine grained natural language inference or classification questions. While it is possible that more appropriate models would perform more competently on the yes/no and counting tasks, the motivation behind our dataset is the evaluation of a model's capacity to comprehensively reason about graph relations that is invariant to specific modes of querying the model.

5 CONCLUSIONS AND FUTURE WORK

We introduced a novel synthetic graph-structured diagram VQA dataset, BizGraphQA, consisting of 10,000 images and over 130,000 question instances, to the information retrieval community. Compared with other comparable datasets, BizGraphQA offers multiple advantages and novelties. Its diagrams are created based on properties from real business graphs across four business domains and are realistically rendered to mimic real documents. BizGraphQA emphasizes challenging inference and reasoning questions including spatial and numeric reasoning and multi-span counting. The mediocre performance of the multi-modal baseline model demonstrates that the dataset raises the VQA challenge to a higher standard. We enthusiastically welcome researchers to explore this new dataset.

We see a number of future directions to pursue with our dataset. First and foremost, we plan to evaluate a larger set of SotA models, including OCR-free ones, such as Donut [11] as well as multimodal large language models like GPT-4 [22], in zero shot and few-shot settings. Second, we aim to expand the inventory of probing questions both in terms of their logic and linguistic diversity. Some of the ways this could be achieved includes prompting of large language models. Lastly, the realism of our diagrams could be potentially enhanced through the end-to-end use of generative models, such as conditional GAN [21] and probabilistic diffusion models [4].

ACKNOWLEDGMENTS

We thank the four anonymous reviewers for their constructive feedback that helped us greatly in improving the paper.

Disclaimer. This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

[1] Brian Davis, Bryan Morse, Brian Price, Chris Tensmeyer, and Curtis Wiginton. 2021. Visual FUDGE: Form Understanding via Dynamic Graph Editing. *arXiv preprint arXiv:2105.08194* (2021).

[2] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation* 32, 5 (05 2020), 829–864. https://doi.org/10.1162/neco_a_01273 arXiv:https://direct.mit.edu/neco/article-pdf/32/5/829/1865303/neco_a_01273.pdf

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

[5] D. A. Hudson and C. D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 6693–6702. https://doi.org/10.1109/CVPR.2019.00686

[6] Drew A. Hudson and Christopher D. Manning. 2019. Learning by Abstraction: The Neural State Machine.. In *NeurIPS*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett (Eds.). 5901–5914. http://dblp.uni-trier.de/db/conf/nips/nips2019.html#HudsonM19

[7] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] While ANLS scores for these domains are superficially higher than the exact match rate, those are expected to be high in general, and thus scores in 60 range should be largely regarded as spurious token matches, especially when it comes to company names.

[8] Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. DVQA: Understanding Data Visualizations via Question Answering. In *CVPR*.

[9] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. FigureQA: An Annotated Figure Dataset for Visual Reasoning. *ArXiv* abs/1710.07300 (2017).

[10] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A Diagram is Worth a Dozen Images. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 235–251.

[11] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free Document Understanding Transformer. arXiv:2111.15664 [cs.LG]

[12] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5583–5594. https://proceedings.mlr.press/v139/kim21k.html

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332

[15] Chen-Yu Lee, Chun-Liang Li, Chu Wang, Renshen Wang, Yasuhisa Fujii, Siyang Qin, Ashok Popat, and Tomas Pfister. 2021. ROPE: Reading Order Equivariant Positional Encoding for Graph-based Document Information Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 314–321. https://doi.org/10.18653/v1/2021.acl-short.41

[16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 121–137.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[18] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

[19] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2582–2591. https://doi.org/10.1109/WACV51458.2022.00264

[20] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2200–2209.

[21] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs.LG]

[22] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[23] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.06434

[24] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[25] Natraj Raman, Sameena Shah, and Manuela Veloso. 2022. Synthetic document generator for annotation-free layout recognition. *Pattern Recognition* 128 (2022), 108660.

[26] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5100–5111. https://doi.org/10.18653/v1/D19-1514

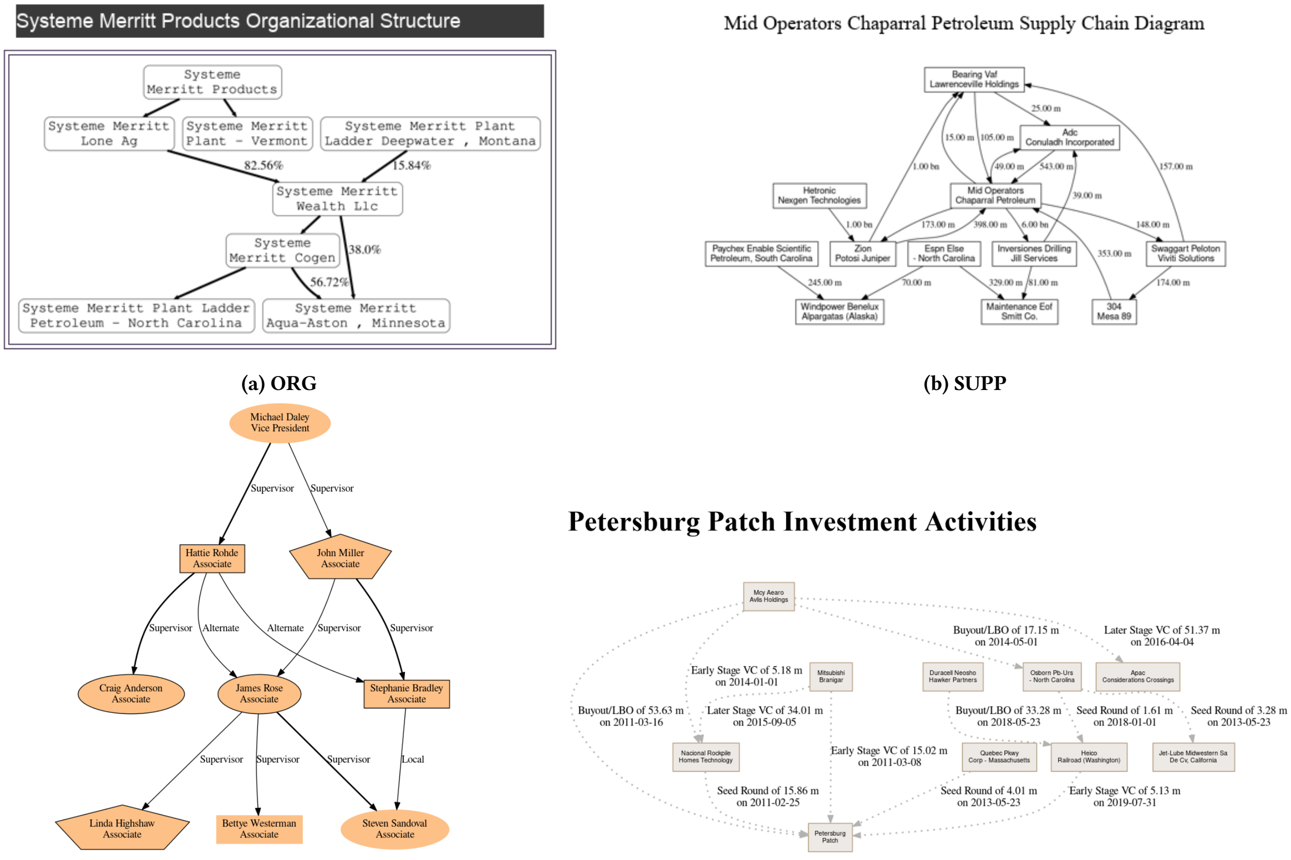
[27] Simon Tanner, Marcelo Feighelstein, Jasmina Bogojeska, Joseph Shtok, Assef Arbelle, Peter Staar, Anika Schumann, Jonas Kuhn, and Leonid Karlinsky. 2022. FlowchartQA: The First Large-Scale Benchmark for Reasoning over Flowcharts. In *Proceedings of DI 2022: The 3rd Workshop on Document Intelligence*. KDD, Washington, DC.

[28] Terry Winograd et al. 1972. Shrdlu: A system for dialog.

[29] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020).

[30] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201

A SAMPLE DIAGRAMS AND QUESTIONS


Systeme Merritt Products Organizational Structure
Systeme Merritt Products
Systeme Merritt Lone Ag
Systeme Merritt Plant - Vermont
Systeme Merritt Plant Ladder Deepwater , Montana
82.56%
15.84%
Systeme Merritt Wealth Llc
Systeme Merritt Cogen
38.0%
56.72%
Systeme Merritt Plant Ladder Petroleum - North Carolina
Systeme Merritt Aqua-Aston , Minnesota
Mid Operators Chaparral Petroleum Supply Chain Diagram
Bearing Vaf Lawrenceville Holdings
25.00 m
15.00 m
105.00 m
Adc Conuladh Incorporated
1.00 bn
49.00 m
543.00 m
157.00 m
Hetronic Nexgen Technologies
Mid Operators Chaparral Petroleum
39.00 m
1.00 bn
173.00 m
398.00 m
6.00 bn
148.00 m
Paychex Enable Scientific Petroleum, South Carolina
Zion Potosi Juniper
Espn Else - North Carolina
Inversiones Drilling Jill Services
353.00 m
Swaggart Peloton Viviti Solutions
245.00 m
70.00 m
329.00 m
81.00 m
174.00 m
Windpower Benelux Alpargatas (Alaska)
Maintenance Eof Smitt Co.
304 Mesa 89

(a) ORG

(b) SUPP

Michael Daley Vice President
Supervisor
Supervisor
Hattie Rohde Associate
John Miller Associate
Supervisor
Alternate
Alternate
Supervisor
Supervisor
Craig Anderson Associate
James Rose Associate
Stephanie Bradley Associate
Supervisor
Supervisor
Supervisor
Local
Linda Highshaw Associate
Bettye Westerman Associate
Steven Sandoval Associate
Petersburg Patch Investment Activities
Mcy Aearo Avlis Holdings
Buyout/LBO of 17.15 m on 2014-05-01
Later Stage VC of 51.37 m on 2016-04-04
Early Stage VC of 5.18 m on 2014-01-01
Mitsubishi Branigar
Duracell Neosho Hawker Partners
Osborn Pb-Urs - North Carolina
Apac Considerations Crossings
Buyout/LBO of 53.63 m on 2011-03-16
Later Stage VC of 34.01 m on 2015-09-05
Buyout/LBO of 33.28 m on 2018-05-23
Seed Round of 1.61 m on 2018-01-01
Seed Round of 3.28 m on 2013-05-23
Nacional Rockpile Homes Technology
Early Stage VC of 15.02 m on 2011-03-08
Quebec Pkwy Corp - Massachusetts
Heico Railroad (Washington)
Jet-Lube Midwestern Sa De Cv, California
Seed Round of 15.86 m on 2011-02-25
Seed Round of 4.01 m on 2013-05-23
Early Stage VC of 5.13 m on 2019-07-31
Petersburg Patch

(c) MGMT

(d) INVST

Figure 6: Sample diagrams from BizGraphQA (cropped from enclosing documents to save page space).

Domain	Type	Sample Question	Sample Answer
ORG	Single Span	Who is the ultimate parent of "Systeme Merritt Lone Ag"?	"Systeme Merritt Products"
	Multi-span	Who are the immediate children of "Systeme Merritt Cogen"?	["Systeme Merritt Aqua-Aston , Minnesota", "Systeme Merritt Plant Ladder Petroleum - North Carolina"]
	Yes/No	Does "Systeme Merritt Aqua-Aston , Minnesota" own 85% or more of "Systeme Merritt Cogen"? Do "Systeme Merritt Wealth Llc" and "Systeme Merritt Plant - Vermont" share the same owner?	False True
	Counting	How many intermediate parents exist between "Systeme Merritt Aqua-Aston , Minnesota" and "Systeme Merritt Products"? Of the immediate children of "Systeme Merritt Plant - Vermont", how many are based in Rhode Island?	3 0
MGMT	Single Span	Who does Craig Anderson report to? What is Hattie Rohde's title?	Hattie Rohde Associate
	Counting	How many managers does Hattie Rohde report to? How many manager relationships are shown here?	1 11
INVST	Single Span	Who was Louisville 1-A's earliest investor? When did Osborn Pb-Urs - North Carolina get its first investment?	Louisville Ales 5/1/2014
	Multi-span	What companies invested in Heico Railroad (Washington)?	[Duracell Neosho Hawker Partners, Osborn Pb-Urs - North Carolina]
	Yes/No	Does S.R.L Csfr invest more in Iii-Gp Parques Nautica Services than in Theatre Connectors Corporation? Does Hornbeck Motoren Lasmo Industries, Alaska receive funds from Covington Szr?	True False
SUPP	Single Span	Who is Windpower Benelux Alpargatas (Alaska)'s top supplier?	"Espn Else - North Carolina"
	Yes/No	Are Swaggart Peloton Viviti Solutions and Zion Potosi Juniper mutual suppliers?	False
	Counting	How many meta-customers (customers of customers) does Hetronic Nexgen Technologies have?	2

Table 4: A subset of BizGraphQA questions broken down by domain and answer type.