Predicting Edges in the Iron March Dataset Using Graph Neural Networks

Alex Newhouse
Georgia Institute of Technology
alex.newhouse@gatech.edu

Christopher Roth
croth37@gatech.edu

David Wu
dpwu@gatech.edu

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length. The abstract section should contain a brief summary of your work that includes the problem statement, proposed solution and results.

1. Introduction

In 2019, an anonymous account published the entire database of the Iron March far-right extremist web forum on archive.org. This database gave researchers unprecedented access to the internal discussions, organizing strategies, and ideological development of some of the most dangerous terrorist networks active in the West. Since the leak, analysts have dug through the data to investigate extreme posting behaviors and community creation, and law enforcement have leveraged it in cases against actual and would-be terrorists.

With hundreds of thousands of forum posts and direct messages, Iron March also represents a uniquely rich dataset for exploring algorithmic methods for extremism research and counterextremism. As an insular social network, the website's data contains information about how far-right extremists connect with one another and organize for real-world action, both as loose political movements and as discrete terrorist cells. While extensive qualitative work has revealed a trove of significant findings about the social network aspect of Iron March, there has been little research done on the application of automated, predictive methods to the data. Successful creation and deployment of automated relationship prediction methods could significantly increase the capacity of researchers, policymakers, and law enforcement to understand and predict how extreme social networks develop. This would likely prove to be a significant boost in efforts to disrupt and mitigate such social networks.

In this paper, we investigate the use of graph neural networks (GNN) to learn social structures from the Iron March data and apply those learnings to predict the creation of new relationships. We explore the impact of node attributes and different GNN architectures, and we find that GNNs show strong potential for use in the extremism research domain. We also propose several areas where future work can build on these findings.

2. Previous Work

In recent years, GNNs have proven capable of achieving high performance across a number of graph-related tasks, from edge and node classification to edge prediction. In this paper, we focus on the use of deep learning methods for edge prediction, while incorporating some findings from other tasks as well.

Edge (or link) prediction has long been reliant on statistical metrics that may not pick up on many nuances in structure. Early machine learning-based edge prediction work relied on similarity metrics, as described in Liben-Nowell and Kleinberg [**?**]. In 2017, Schlichtkrull *et al*. [**?**] pioneered the use of graph convolutional networks (GCN) for the edge prediction task, building on the GCN framework proposed by Duvenaud *et al*. [**?**] and establishing the practice of autoencoding graph data. With SEAL, Zhang and Chen [**?**] propose extracting local subgraphs as a heuristic in a GCN-based model, achieving new heights of performance on the edge prediction task.

Gu *et al*. [**?**] reached state-of-the-art results on link prediction tasks via a modified graph attention network (GAT) called DeepLinker, which demonstrated improvements in computational time, prediction accuracy, and automated discovery of node importance metrics. Further, Izadi *et al*. [**?**] successfully beat the state-of-the-art score on the CORA node classification task via the inclusion of natural gradient descent in the graph neural network's optimization problem.

Zhang *et al*. [**?**] describe a method for adding additional knowledge into a graph to assist with link prediction. The

authors analyzed a citation graph and used named entity recognition to add additional knowledge edges to the graph. This combats the sparsity problem that is common in most real world graph data. Further extending the idea of mining text for use in supplementing graph data, Zhang *et al*. [**?**] used a bag of words approach to construct node features.

3. Dataset

The Iron March database is real-world data: it was generated as a byproduct of people interacting with the website and one another, and it was not designed for the purpose of training machine learning models. In order to train the GNN models, we use the essential data in the database: forum posts and direct messages. These represent the thoughts and connections of Iron March's users over the entire course of the website's existence from 2011-2017.

We train and validate the models on the set of forum posts. These posts are "threaded" in a similar structure to Reddit or 4chan posts, which allows for the generation of a social network based on shared interactions within threads. There are over 195,000 forum posts in total, of which roughly 182,000 have an identifiable author and are thus usable. Forum posts consist of a number of features, but for the purposes of this paper we focus on three: the raw text of the post, the author ID, and the thread ID.

In addition, we use Iron March's direct message data in order to further explore the capabilities of the model. There are about 22,300 direct messages, of which roughly 21,000 have an identifiable author. Like forum posts, messages are identified by a thread ID, facilitating the creation of an edgelist.

The full Iron March dataset contains sensitive data, including political beliefs, ages, names, locations, email addresses, and other social account data. The data was leaked without consent of the data subjects, further adding to its sensitivity. However, the dataset has been extensively published and covered in both reputable news sources and academic writing, and it has been used in government documents for the purpose of justifying indictments. Many of its subjects have been widely identified in the media.

We carefully use this data by not including personal information (beyond that which may incidentally arise in forum or message posts) in training the models. In addition, we do not identify any subject in subsequent analysis who has not already been identified in the public record, either through arrest warrants or coverage in reputable sources.

4. Approach

We applied findings from recent edge prediction studies to the novel problem of predicting Iron March edges. Specifically, we explored the effectiveness of graph convolutional neural networks. Although developed primarily for use in analysis of knowledge and citation graphs, we hypothesized that social networks should share a similar enough structure to work effectively in a GNN.

Iron March data does not contain an intrinsic individual-to-individual relationship. There are not friends lists as there are on e.g. Facebook, nor follower/followee relationships as there are on e.g. Twitter. Rather, all relationships are latent within interactions on the forums or in direct messages. As such, we construct an individual-to-individual edgelist via first building a two-mode or bipartite network between individuals and forum threads. From there, we extract individual-to-individual connections if two users post in the same forum thread.

Since this heuristic for generating relationships generally exaggerates the connections–two individuals who happen to post on the same forum thread only once likely are not close friends–we also implement a threshold for inclusion in our edgelist. In this case, we implement a threshold of three: an edge is created between two individuals if they post on three of the same forum threads.

This edgelist is used for training a GNN. Leveraging Pytorch Geometric's built-in dataset splitting functionality, we create an 85%, 10%, 5% split for train, validation, and test data, respectively. We implement the training functionality detailed in Zhang and Chen [**?**], wherein a number of unconnected node-pairs are selected to create a "negative" edgelist, and concatenated onto the "positive" edgelist. The model can then treat these as the negative and positive labels for the prediction task.

The implementation of this GNN relies on the code in Pytorch Geometric's link prediction example, from Fey and Kim [**?**]. We customized this code with the functionality necessary to process the Iron March data, generate bag-of-words features based on the text posts, and pass it through the GNN. As part of experimentation, we significantly varied the architecture of the model, ultimately selecting a much deeper and larger model than was used in the original repository. The metrics were modified to output accuracy in addition to AUC and loss, and we included code to generate loss and accuracy curves. Finally, we added functionality to run inference on the entire network after training, so that we can compare the existing network to the predicted network.

4.1. Bag of Words and Other Node Features

The first node feature we used was bag-of-words. This is one method to convert text posts into values for use in the GNN. We first took all the posts and created a vocabulary of all the words. Then, for each author on the forum, we took their posts and counted the frequency of each word they used from the vocabulary, giving us a vector of vocabulary length for each author where the frequency of the i-th vocabulary word used by the author is held at the i-th position. The resulting bag-of-words is quite large, being of

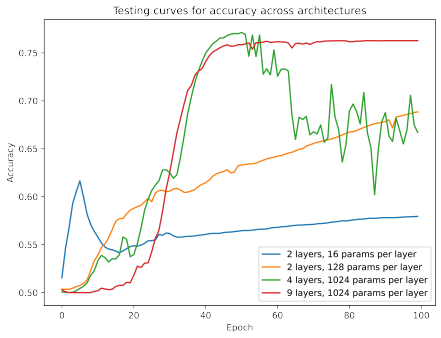

Figure 1. Comparison of accuracy curves for different model architectures.

shape (number of authors) x (number of words in vocabulary).

To trim down the size of this and make sure the vocabulary contains valid words, the forum posts are pre-processed by removing HTML tags, removing punctuation, and removing stop words (the most common words in a language).

5. Experiments and Results

5.1. Model Architectures

Our first attempt to train a GNN on this data utilized the existing, two-layer graph convolutional autoencoding network that is default in the Pytorch Geometric repository. While early experiments showed promise on the smaller direct message dataset, the main forum dataset proved too large for effective training on this simple network. While the loss metric on the training dataset indicated that the model was learning something, the scores on the validation and test datasets revealed that it was not generalizing at all.

As a result, our experimentation largely focused on adding more complexity to the model. The straightforward method of adding more GCN layers, and increasing the parameters at each layer, proved to be effective at learning a model that was able to perform better on the train and test sets. We achieved AUC scores on both sets near 1.0, while accuracy scores ranged from a little over 50% on smaller models to roughly 75% on eight-layer models with 1024 parameters per layer. The randomness of the train/test split in the data resulted in a high variance of accuracy curves, especially in smaller models, so we averaged accuracy scores over 10 runs to compare across architectures.

In order to sanity-check our model, we produced an inference edgelist by passing the entire graph through the best-performing model and comparing it to the original edgelist. The graph for each is shown below.

[Network graphs here]

(10 points) How did you measure success? What experiments were used? What were the results, both quantitative and qualitative? Did you succeed? Did you fail? Why? Justify your reasons with arguments supported by evidence and data.

Important: This section should be rigorous and thorough. Present detailed information about decision you made, why you made them, and any evidence/experimentation to back them up. This is especially true if you leveraged existing architectures, pre-trained models, and code (i.e. do not just show results of fine-tuning a pre-trained model without any analysis, claims/evidence, and conclusions, as that tends to not make a strong project).

6. Other Sections

You are welcome to introduce additional sections or subsections, if required, to address the following questions in detail.

(5 points) Appropriate use of figures / tables / visualizations. Are the ideas presented with appropriate illustration? Are the results presented clearly; are the important differences illustrated?

(5 points) Overall clarity. Is the manuscript self-contained? Can a peer who has also taken Deep Learning understand all of the points addressed above? Is sufficient detail provided?

(5 points) Finally, points will be distributed based on your understanding of how your project relates to Deep Learning. Here are some questions to think about:

What was the structure of your problem? How did the structure of your model reflect the structure of your problem?

What parts of your model had learned parameters (e.g., convolution layers) and what parts did not (e.g., post-processing classifier probabilities into decisions)?

What representations of input and output did the neural network expect? How was the data pre/post-processed? What was the loss function?

Did the model overfit? How well did the approach generalize?

What hyperparameters did the model have? How were they chosen? How did they affect performance? What optimizer was used?

What Deep Learning framework did you use?

What existing code or models did you start with and what did those starting points provide?

Briefly discuss potential future work that the research community could focus on to make improvements in the direction of your project's topic.

Student Name	Contributed Aspects	Details
Team Member 1	Data Creation and Implementation	Scraped the dataset for this project and trained the CNN of the encoder. Implemented attention mechanism to improve results.
Team Member 2	Implementation and Analysis	Trained the LSTM of the encoder and analyzed the results. Analyzed effect of number of nodes in hidden state. Implemented Convolutional LSTM.

Table 1. Contributions of team members.

7. Work Division

Please add a section on the delegation of work among team members at the end of the report, in the form of a table and paragraph description. This and references do **NOT** count towards your page limit. An example has been provided in Table 1.

8. Miscellaneous Information

The rest of the information in this format template has been adapted from CVPR 2020 and provides guidelines on the lower-level specifications regarding the paper's format.

8.1. Language

All manuscripts must be in English.

8.2. Dual submission

Please refer to the author guidelines on the CVPR 2020 web page for a discussion of the policy on dual submissions.

8.3. Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2020.**

Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven.

8.4. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment

the `\cvprfinalcopy` command in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is 095.5), although in most cases one would expect that the approximate location will be adequate.

8.5. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: http://www.pamitc.org/documents/mermin.pdf.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper "further details may be found in [5]". Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al*.

You can handle this paper like any other. Don't write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn't handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

Q: Are acknowledgements OK?
A: No. Leave them for the final copy.

Q: How do I cite my results reported in open challenges?
A: To conform with the double blind review policy, you can report results of other challenge participants together with your results in your paper. For your results, however, you should not identify yourself and should not mention your participation in the challenge. Instead present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison to other results.

8.6. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TEXbook, p165.

The space after *e.g*., meaning "for example", should not be a sentence-ending space. So *e.g*. is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al*." (not "*et. al.*" as

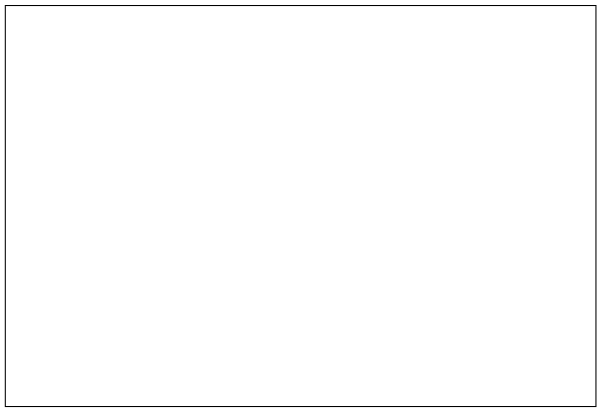

Figure 2. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

"*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [1], and subsequently developed by Alpher and Fotheringham-Smythe [2], and Alpher *et al*. [3]."

This is incorrect: "... subsequently developed by Alpher *et al*. [2] ..." because reference [2] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [2, 1, 4] to [1, 2, 4].

8.7. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

8.8. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high.

8.9. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 2. Results. Ours is better.

Figure and table captions should be 9-point Roman type as in Figures 2 and 3. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

8.10. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

8.11. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [4]. Where appropriate, include the name(s) of editors of referenced books.

8.12. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

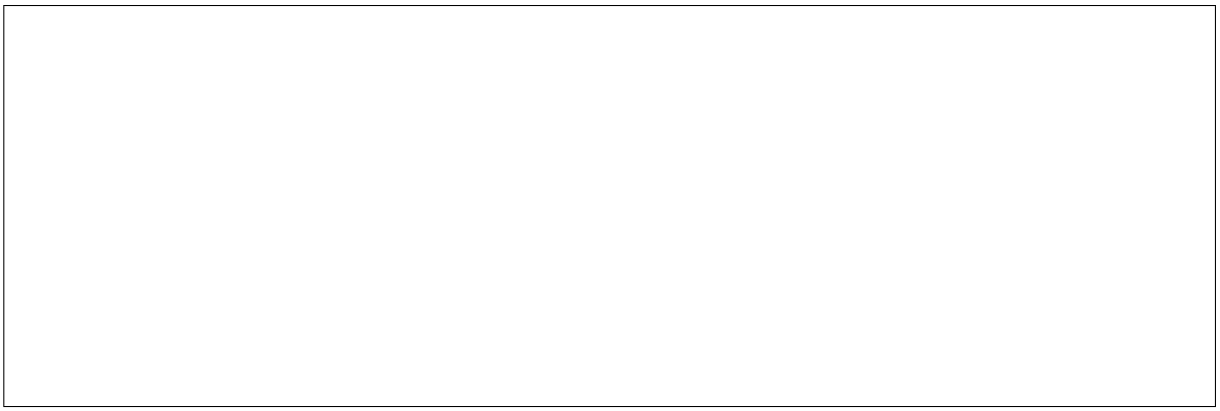

Figure 3. Example of a short caption, which should be centered.

not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

8.13. Color

Please refer to the author guidelines on the CVPR 2020 web page for a discussion of the use of color in your document.

References

[1] Link prediction with graph neural networks and knowledge extraction, 2020. 1, 2

[2] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1

[3] Matthias Fey and Dongkwan Kim. Pytorch geometric - link prediction. https://github.com/rusty1s/pytorch_geometric/blob/master/examples/link_pred.py, 2021. 2

[4] Weiwei Gu, Fei Gao, Xiaodan Lou, and Jiang Zhang. Link prediction via graph attention network. *CoRR*, abs/1910.04807, 2019. 1

[5] Mohammad Rasool Izadi, Yihao Fang, Robert Stevenson, and Lizhen Lin. Optimization of graph neural networks with natural gradient descent. *CoRR*, abs/2008.09624, 2020. 1

[6] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031, May 2007. 1

[7] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017. 1

[8] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1, 2