Reviewer #1

Q1: This paper is not clear as there are some typos in the text and the methodology section 4 is not well explained.

A1: Thank you very much for pointing out this issue. We will carefully revise and re-organize our paper to make it clearer and easier to understand. Here, we will clarify potential misunderstanding, if any, for our methods. For the methodology section (Section 4), we would try our best to offer additionally plain language explanations for the high-level idea of our proposed methods. However, as our motivation for deriving consistent methods based on the data generation model is purely theoretical, we need to give formal mathematical definition and derivations, which are important, necessary and may not be simply replaced by an easy-to-understand plain language explanation.

Q2: While the authors present prior works in their paper, they do not do a good job of explaining how these works relate or differ from each other. They do motivate their work with limitations of existing methods.

A2: Thank you for your comments. We agree that we motivated our work from existing methods, i.e., none of the existing methods considers a data generation process and the consistency of these methods would be hardly guaranteed. Our paper is the first work that solves the partial-label learning problem (with two novel provably consistent methods) through the lens of the data generation process. Previous works improve the practical performance of partial-label learning using various strategies, including the EM iterative procedure [17,22,28], maximum margin [31], manifold regularization [32,23], and error-correcting output codes [34]. Theoretical works [18,19] make the same assumption on ambiguity degree, with the difference that [18] provides a classifier-consistent method while [19] only focuses on learning theory. We will take your valuable suggestion to have a related work section that compares these methods and their limitations, and provide more detailed explanations in our final version.

Reviewer #2

Q1: I do not know the difference between the suggested data distribution and a general noisy label setting.

A1: There is an evident difference of the problem setting between *partial-label learning* (PLL) and *noisy-label learning* (NLL). Specifically, we denote by the data distribution for PLL $p(\boldsymbol{x}, Y)$ where Y is a set of candidate labels and the data distribution for NLL $p(\boldsymbol{x}, \tilde{y})$ where $\tilde{y}$ is a single observed label that may not be the correct label. We can see that a set of candidate labels is generated by the former distribution, while a single observed label is generated by the latter distribution. Two different data forms cannot share the same data distribution. In our proposed data distribution for PLL, each possible candidate label set that contains the correct label will be uniformly sampled as the observed candidate label set. Since there are in total $2^{k-1} - 1$ possible candidate label sets, each candidate label set will be chosen to be the observed candidate label set with probability $1/(2^{k-1} - 1)$. For example, suppose the label space is $\{1, 2, 3, 4\}$ (i.e., $k = 4$) and the true label is 2 for a given instance, there are $2^{k-1} - 1 = 7$ possible candidate label sets: $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{2, 3, 4\}$, $\{1, 2\}$, $\{2, 3\}$, $\{2, 4\}$, $\{2\}$, each of them would be selected as the observed candidate label set with probability $1/7$. In contrast, in the data distribution (symmetric noise) for NLL, the correct label for a given instance has some probability (denoted by z) to be the observed label, and each of the other labels has a probability of $(1 - z)/(k - 1)$ to be the observed label. This concrete example also shows the difference between the two data distributions.

Reviewer #3

Q1: I am a bit confused about why we need a classifier-consistent method since it is worse than the risk-consistent method. Is it because some literature focusing on proposing a classifier-consistent method, or there are some limitations on the loss functions used? I would suggest the paper to have more discussion on this part.

A1: Thank you for your insightful comments. Yes, it is partly because some literatures focus on proposing a classifier-consistent method. We would like to show that based on our proposed data generation process, we can also derive a novel classifier-consistent method. Another reason is that in *noisy-label learning*, classifier-consistent methods are generally better than risk-consistent methods because risk-consistent methods usually overfit due to the negative empirical risk issue. In *partial-label learning* (PLL), we would like to provide the first study that compares the two types of consistent methods, which motivates us to also propose a classifier-consistent method. It is interesting that our risk-consistent PLL method does not have the overffiting issue since there is no negative empirical risk, hence it outperforms the classifier-consistent PLL method. In addition, we would like to emphasis the importance of the data generation process. We would like to show that having such data distribution, we are able to derive consistent (risk-consistent and classifier-consistent) methods. Both consistent methods can be considered as the products of the data generation process. We will provide more discussions on the difference between the risk-consistent method and the classifier-consistent method. There is no limitation on the used loss functions for the risk-consistent method. There is a small restriction (i.e., Lemma 3) on the used loss functions for the classifier-consistent method, while we have shown that common loss functions (e.g., the softmax cross entropy loss and mean squared error) can satisfy this condition. Intuitively, the performance of the two methods would be affected by the used loss functions. It would be interesting to investigate the influence of different loss functions on our methods in future work.