We thank each of the reviewers for their comments and suggestions.

Clarification on sparsity regime (raised by R_1 and R_2).

In our submission, we used the upper-bound notation $a, b = O(\log n)$ which includes $O(1)$.

We did so because behaviour exhibited in different regimes is fundamentally dichotomous. We wanted to be succinct in gathering all of the results in different regimes.

As seen from Line 63 for *large changes*, reliable testing is possible in the $O(1)$ sparsity regime - the comments following Thm. 2 explore this. Nevertheless, for *small changes* (see Line 67, and Thm. 1), it is impossible to reliably test with $a, b = O(1)$ - rather, $a, b = \Omega(\log n)$ is both necessary and sufficient for testing small changes. Overall, this is one of the main messages of the paper: the SNR requirements for testing large and small changes are qualitatively different.

Thus, to present all of our testing results, we must technically permit a, b to vary from constant to logarithmic, even though we are primarily focused on what is possible when $a, b = O(1)$. That said, we will revise the surrounding text to provide a bit more context for this technical note.

In addition, our thanks to R_2 for bringing to attention our oversight in not citing Mossel, Neeman, and Sly here, which we will amend.

Practicality (raised by R_2 and R_3). First, we note that our approach can be naturally extended to more practical settings (e.g., many communities, degree correction). For instance, for multi-community setting, given the number of communities, one can again recover weakly (e.g., by leveraging SDP methods of Fei and Chen), and then adapt our statistics in a straightforward manner. However, this will affect the SNR thresholds for testing, and fully characterising the dependence on the number of communities, etc. is messy and non-trivial, enough to merit further work (mirroring the development of the recovery literature).

Similarly, while we present preliminary exploration of this in the experiments section, validation of these methods and models on real-world networks is a non-trivial task. We think that extensive pursuit of these here would distract from the primary theoretical considerations of the paper.

We thank you each for bringing up these important questions, and we will include a discussion of these as directions for future research.

Constants (raised by R_2). Note that, as stated on line 157, each constant in the paper can be *explicitly bounded*, although these bounds may not be the tightest possible. For instance, in the limit as $n \nearrow \infty$, the TST result for large changes shows impossibility of reliable testing if $\Lambda < 1$, and also that $\Lambda > 4$ is sufficient for reliable testing. The lower bound is explicitly discussed in the proofs, see Line 866 of the supplement. The upper bound follows from the proof of Thm. 2 and the work of Mossel, Neeman, and Sly. Non-asymptotic results are also discussed.

From our perspective such gaps within a constant factor of each other is acceptable. Establishing exact constants in the SBM can be quite challenging, and, even for recovery, these are only known in very particular settings, such as exact recovery or weak recovery with distortion $n(1/2 - \varepsilon)$. This is certainly an interesting problem, one that likely requires a dedicated effort to resolve. We will add a discussion of this question to a future work section.

Balance (raised by R_2). We will try to highlight the balance assumption in the introduction, although we note the use of 'balance' in the title precisely for this purpose. (As an aside, since submission we have extended the theorems to unbalanced but linearly-sized communities, perhaps ameliorating this concern.)

Lines 183,184 (raised by R_1). This was meant to be a brief comment explaining why we switch, based on the relative sizes of a and b, from a test based on counts of edges across communities $N_a^{x_0}(G)$ to one based on counts of edges within communities $N_w^{x_0}(G)$. Both of these counts are of roughly the same signal strength, but depending on the regime ($a > b$ vs. $a \leq b$), one count will have a lower noise level. We will clarify this comment so that it reads more smoothly.

Typographical errors, and presentation suggestions. We thank the reviewers for pointing these out, especially R_1 for catching one in a definition (the $\sup$ should be over $d = 0$ *or* $d \geq s$)! We will correct these.